НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Факультет інформатики та обчислювальної техніки
Кафедра обчислювальної техніки

До захисту допущено:

Завідувач кафедри
_____ Сергій СТІРЕНКО

“___”_____2020 р.

Дипломний проект

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Комп’ютерні системи та мережі»

спеціальності 123 «Комп’ютерна інженерія»

на тему: «Програмна система QSAR моделювання для оцінки здатності блокування реплікації ВІЛ»

Виконав:
студент IV курсу, групи ІО-63
Антон КЕЛЕБЕРДА _____

Керівник:
Професор, д.т.н.,
Михайло НОВОТАРСЬКИЙ _____

Консультант з нормоконтролю:
Професор, д.т.н.,
Валерій СІМОНЕНКО _____

Рецензент:
Доцент, к.т.н.
Марія ОРЛОВА

_____

Засвідчую, що у цьому дипломному проекті немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2020 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Факультет інформатики та обчислювальної техніки
Кафедра обчислювальної техніки

Рівень вищої освіти – перший (бакалаврьский)

Спеціальність – 123 «Комп'ютерна інженерія»

Освітньо-професійна програма «Комп'ютерні системи та мережі»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Сергій СТІРЕНКО

«___» _____2020 р.

ЗАВДАННЯ
на дипломний проект студента

Келеберди Антона Миколайовича

1. Тема проекту «Програмна система QSAR моделювання для оцінки здатності блокування реплікації ВІЛ»
керівник проекту Новотарський Михайло Анатолійович, д.т.н., професор
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом по університету від «07» травня 2020 року №1081-с
2. Термін здачі студентом закінченого проекту
_____
3. Вихідні дані до проекту *технічна документація*
4. Зміст розрахунково-пояснювальної записки (перелік питань, які розробляються)
Опис та аналіз предметної області, аналіз аналогів у даній області, дослідження основних вимог до ПЗ, конструювання архітектури системи та створення компонентів системи .
5. Перелік графічного матеріалу
Принципова схема, функціональна схема та структурна схема

6. Консультанти проекту, з вказівкою розділів роботи, які до них вносяться

Розділ	Консультант	Підпис, дата	
		Завдання видав	Завдання прийняв
нормоконтроль	д.т.н., проф. Сімоненко В.П.		

7. Дата видачі завдання_____

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Найменування етапів дипломного проекту	Строк виконання етапів проекту	Примітки
1.	*Затвердження теми роботи*	*1.09.2019*	
2.	*Вивчення та аналіз завдання*	*2.09.2019-11.03.2020*	
3.	*Розробка архітектури та загальної структури програми*	*11.03.2020-22.03.2020*	
4.	*Розробка структур окремих Інтерфейсів програми*	*22.03.2020-2.04.2020*	
5.	*Програмна реалізація*	*2.04.2020-15.04.2020*	
6.	*Оформлення пояснювальної записки*	*15.04.2020-21.05.2020*	
7.	*Захист програмного продукту*	*21.05.2020 – 25.05.2020*	
8.	*Передзахист*	*26.05.2020*	
9.	*Захист*		

Студент Антон КЕЛЕБЕРДА

Керівник Михайло НОВОТАРСЬКИЙ

ВІДОМІСТЬ ДИПЛОМНОГО ПРОЕКТУ

№ з/п	Формат	Позначення	Найменування	Кількість листів	Примітка
1.	А4		Завдання на дипломний проект	2	
2.	А4	ІАЛЦ.466500.002 ТЗ	Технічне завдання	3	
3.	А4	ІАЛЦ.466500.003 ПЗ	Пояснювальна записка	54	
4.	А4	ІАЛЦ.466500.004 А1	Принципова схема алгоритму	1	
5.	А4	ІАЛЦ.466500.005 А2	Функціональна схема	1	
6.	А4	ІАЛЦ.466500.006 А3	Структурна схема	1	

АНОТАЦІЯ

Робота присвячена розробці програми для прогнозування здатності різних хімічних сполук блокувати реплікацію вірусу імунодефіциту людини, а також для оцінки побічних дій препаратів. У зв'язку з актуальністю проблеми поширення ВІЛ/СНІДу виникає потреба у пошуку профілактичних заходів для запобігання цієї проблеми.

Запропонований метод полягає у пошуку хімічних сполук, здатних повністю або частково блокувати реплікацію вірусу імунодефіциту людини в організмі.

ABSTRACT

The work is devoted to the development of a program for predicting the ability of various chemical compounds to block the replication of human immunodeficiency virus, and also for assessing the side effects of drugs. Due to the urgency of the problem of the spread of HIV/AIDS, there is a need to seek preventive measures to prevent this problem.

The proposed method is to search for chemical compounds that can completely or partially block the replication of human immunodeficiency virus in the body.

Технічне завдання до дипломного проекту

ЗМІСТ

1. НАЙМЕНУВАННЯ ТА ОБЛАСТЬ ЗАСТОСУВАННЯ

Дане технічне завдання розповсюджується на розробку програмної системи QSAR моделювання для оцінки здатності блокування реплікації ВІЛ та для оцінки побічних дій препаратів.

Область застосування: QSAR моделювання застосовується для прогнозування біологічної активності хімічних сполук та їх фізико-хімічних властивостей, таких як температура кипіння, здатність до розчинення у воді та ліпідах, константи кислотності та основності.

2. ПІДСТАВИ ДЛЯ РОЗРОБКИ

Підставою для розробки служить завдання на виконання розробки програмної системи QSAR моделювання для оцінки здатності блокування реплікації ВІЛ та для оцінки побічних дій препаратів, затвердженою кафедрою обчислювальної техніки Національного технічного Університету України «Київський Політехнічний Інститут імені Ігоря Сікорського».

3. МЕТА ТА ПРИЗНАЧЕННЯ РОЗРОБКИ

Метою даного проекту є розробка програмної системи QSAR моделювання для оцінки здатності блокування реплікації ВІЛ та для оцінки побічних дій препаратів.

4. ДЖЕРЕЛА РОЗРОБКИ

Джерелами для розробки служать науково-технічна література, публікації в періодичних виданнях, довідники та публікації в Інтернеті за даним питанням.

5. ТЕХНІЧНІ ВИМОГИ

5.1. Вимоги до розроблюваного продукту

- Розробка графічного інтерфейсу;
- Розробка програми для QSAR моделювання та прогнозування властивостей хімічних сполук;
- Розробка засобів візуалізації результатів моделювання;

5.2. Вимоги до програмного забезпечення

- Операційна система Windows 7, MS Windows 8/8.1, MS Windows 10
- Інтерпретатор Python 3.5 і вище

5.3 Вимоги до апаратного забезпеення

- Комп’ютер на базі процесора Intel Core i3 і вище
- Оперативної пам’яті не менше 4 ГБ

ПОЯСНЮВАЛЬНА ЗАПИСКА

до дипломного проекту

на тему: «Програмна система QSAR моделювання для оцінки здатності блокування реплікації ВІЛ»

Київ – 2020

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ВІЛ - Вірус Імунодифіциту Людини

QSAR - Пошук кількісних співвідношень структура-властивість

SMILES - Специфікація спрощеного представлення молекул в рядку введення — система правил

PASS - прогноз спектральної біологічної активності органічних сполучень

DAPY – діарилпіримідини (найкращий засіб для боротьби з ВІЛ)

ВСТУП

На сьогодні дедалі більшої актуальності набуває проблема поширення ВІЛ-інфекції та проблема пошуку методів її запобігання. Використовуючи технології машинного навчання можна знайти хімічні сполуки, які здатні блокувати реплікацію вірусу. Пошук таких сполук виконується на основі наборів даних з відомими сполуками та їх здатністю до блокування реплікації.

Предметом розгляду в роботі є створення програмної системи QSAR моделювання для оцінки здатності блокування реплікації ВІЛ та для оцінки побічних дій препаратів.

Для цього необхідно вирішення наступних завдань:

- Проведення огляду
- Наведення основних відомостей про проект:
    - Аналіз і характеристика об’єкту проектування
    - Обґрунтування оптимального варіанта реалізації програми
- Визначення вимог до програмного забезпечення (Технічне завдання на проектування)
- Розробка і опис основних функціональних можливостей
- Проектування програми для обміну повідомленнями на ОС Android (Опис алгоритму і програмного забезпечення)
- Розробка програмної системи:
    - Вибір і обґрунтування структури проектування системи
    - Основні рішення з реалізації програми в цілому і її компонентів

РОЗДІЛ 1
ОГЛЯД ІСНУЮЧИХ РІШЕНЬ (ПРОГРАМА PASS)

В наш час переважає напрямлений підхід до пошуку і створення нових лікарських препаратів: хімічні сполуки перевіряються на невелику кількість видів біологічної активності, а властивості виявлених структур оптимізуються шляхом синтезу і дослідження їх аналогів. При цьому більшість видів біологічної активності, присутні у досліджуваній речовині, але являються побічними відносно напрямку вивчення, залишаються невивченими. Кожна речовина може проявляти декілька видів біологічної активності. Деякі з них знаходять токсичні побічні ефекти, інші стають основою для реєстрації лікарського препарату по новому напрямку. Так наприклад ацестазолон був запропонований в якості діуретика в 1954 році, і як проти епілептичний засіб в 1956 році.

Таким образом, існує певна протилежність між направленістю вивчення біологічних активних сполучень і множинністю біологічних ефектів, які потенційно може показати кожна речовина. [1]

Біологічна активність є результатом взаємодії речовини з біологічним об'єктом. Вона залежить від характеристики речовини (структури його молекул і фізико-хімічних властивостей), біологічного об'єкта (вид, стать, вік) і способу взаємодії (місце введення, доза). Весь комплекс біологічних ефектів, які речовини можуть викликати при деяких умовах взаємодії з біологічними об'єктами, без врахування особливості конкретного експерименту, будем називати спектром біологічної активності речовини. Це якісна характеристика речовини, залежна тільки від структури його молекули. З іншої сторони, великий масив даних можна зібрати, тільки використовуючи багато різних джерел, оскільки інформація в одної публікації ніколи не описує всі аспектів біологічної дії описаної в речовині. Наприклад кофеїн являється стимулятором, аналептиком і діуретиком.

Практично неможливо дослідити експериментально всі відомі види активності ні одної хімічно сполуки. Навіть якщо прийняти до уваги можливості сучасного комплексу досліджень, який проводиться по відношенню одного або декількох досліджень лікарських препаратів, розглядається в перспективі в конкретних період часу. Реальну можливість комплексного вивчення біологічної активності речовини

може забезпечити новими технологіями комп'ютерного прогнозування. Більшість існуючих в наш час комп'ютерних методів молекулярного моделювання і аналізу зв'язків "структура-активність" використовувати для дослідження взаємодії і оптимізації властивостей базових структур на основі аналізу кількісних співвідношень "структура-активність" в рамках одного хімічного класу. Методи молекулярних спів падань і кластеризації, також можуть використовуватись для розділення сполучень на структурні групи, прогнозу біологічної активності і відбору речовин. Реальні методи хімічної інформатики описані в чотирьох томах, виданих за редакції Йохана Гастайгера "Handbooks of Cheminformatics From Data to Knowledge ".

Запобігання розриву між одиничними активностями, які беруть до уваги на кожній стадії дослідження, і багато видів біологічної активності органічних активності з допомогою комп'ютерної системи. Спроби створити подібну схему багаторазово проводились раніше. Можливості комп'ютерного прогнозування біологічної активності проводилось в СССР в рамках державної системи реєстрації всіх синтезованих хімічних сполучень. Ця задача не була вирішена, по ряду причин, але була створена база для створення комп'ютерної системи в наш час.

Нами розглянута комп'ютерна система PASS (Prediction of Activity Spectral Substrances – прогноз спектральної біологічної активності органічних сполучень), основана на аналізі взаємодій "структура-активність" з використанням навчальної вибірки, містять велику різновидність хімічних сполук з різними видами біологічної активності. Точність прогнозу залежить від різних факторів, але в справжній час важливим фактором є якість вибірки. Хороша вибірка має включати біологічно активні речовини з кінцевою кількістю інформації. Для кожної речовини із вибірки, має бути вивчений весь спектр біологічної активності, але немає достатньо просторої бази даних хімічних сполук, які дослідили на всі види активності, неповнота інформації в біологочній активності є у всіх базах, а повне експериментальне дослідження неможливе. Для дослідження всіх відомих видів біологічної активності не вистачить ресурсів всієї біосфери, ні ресурсів людства.

Із перерахованої вище інформації про прогноз біологічної активності хімічних сполук, можна сформулювати наступні вимоги [2]

- Велика кількість і різновидність прогнозованих видів активності

- Інформація в реальній вибірці далека від ідеалу
- Здатність прогнозувати біологічну активність сполучень різних хімічних класів з допустимою точністю
- Дослідження мінімальної інформації про речовини, достатньої для прогнозу спектральної активності нових і ще не синтезованих сполук

В реальний час система PASS постійно розвивається і частково підходить під наші вимоги, а її основними складовими є

- Уява про біологічну активність
- Опис структури хімічних сполук
- База даних про взаємозв'язки "структура-активність"
- Алгоритм прогнозу спектрів біологічної активності

Система PASS постійно розвивається, і актуальна версія 1,917 має ряд значних відмінностей.

1.1 Представлення біологічної активності

Для кожної сполуки є список видів активності, які можуть проявитись при різних умовах. В PASS приймається, що речовина не має видів активності, які не вказані в спектральному аналізі. Хоча неможна виключити ситуації, коли інформація не про активність не була найдена. Це приближення не має сильного впливу на результати аналізу взаємозв'язків "структура-активність" і виконуваного на його основі прогнозу, завдяки статистиці в алгоритмах PASS.

Для прогнозів із допомогою PASS можна використати різні способи класифікації органічних сполук. Якщо класи дійсно визначаються особливостями структури молекул, то прогноз можна вважати цілком успішним. Наприклад інтервал значень деяких величин можна розглядати в PASS, як активність, якщо значення величини належать інтервалу, то речовина активна. Тому існує багато способів застосунку PASS.[3]

1.2 Опис хімічної структури

Результати взаємодії речовини на біологічний об'єкт, при рівних умовах, визначається структурою його молекул. В хімічній практиці речовина описується

структурною формулою і набором фізико-хімічних параметрів. При аналізі взаємозв'язків "структура-активність" використовуються багато різних характеристик хімічних сполучень: структурні фрагменти, геометричні і топологічні індекси, фізико-хімічні параметри. Для різних видів біологічної активності в SAR/QSAR існують різні дескриптори, які описують структуру молекули.

Структурна формула, традиційно записується відповідно до номенклатурних правил в хімії, показує атомний склад і розташування атома в молекулі. Хоча реальна речовина представлена в суміші компонентів різного характеру: молекули, іони в різних конформаційних і електронних станах, таутоміри, комплекси молекул в різних видах і степенях стійкості. Склад суміші залежить від зовнішніх умов. Структурна формула лише характеризує одну компоненту, умовно прийняту за основну. Хоча всі стани молекул пов'язані між собою, а їх взаємодія може бути дуже складною. Тому кожен стан молекули можна використовувати для опису її структури.[4]

В не ідеальних умовах, при взаємодії з іншими молекулами – мішенями, рицепторами, ферментами,- замість координат атомів варто розглядати функції зміни координат атомів. В такій ситуації більш зручним способом для практичного використання структури являється вибірка із відповідного набору координат атомів. Хоча при використанні такого способу описання молекул, потрібні значні ресурси для проведення квантово хімічних обрахунків, і виникає не вирішена проблема перетворення координат повного опису. В такому сенсі можна сказати, що структурна формула однозначно визначає властивості молекул. Тому в PASS в якості основи для опису структури органічних сполучень вибрана структурна формула.

1.3 Бази даних і значення в SAR base

Для прогнозів в PASS використовують **SAR base,** яка створюються на основі аналізу вибірки, з структурними формулами і спектром активності органічних сполучень. **SAR base** містить в собі словник зі видамами біологічної активності, словник MNA- дескриптор, опис структур і активності речовин в виборці, дані і знання про взаємодію "структурно-біологічна активність ".

В версії PASS 1,917 міститься 57978 структур молекул і спектрів активності ліків, і біологічно активних речовин. Словник MNA- дескриптор включає 44041 дескрипторів 1-го і 2- го рівнів.

В різних джерелах інформації про біологічну активність описана різними термінами, тому спектри активності в навчальній вибірці стандартизовані SAR base версії 1,917 загальна кількість різних видів біологічної активності 4463, але 769 з них представлені тільки з одних з'єднанням, 504- двома, а трьома і понад 3190, в список прогнозованих за замовленням видів активності вкладено 2005 назв, з них 224 це фармакологічні ефекти, 1756 молекулярні механізми і 25 різних побічних ефектів. Середня точність прогнозів 88%. [5]

1.4 Алгоритм прогнозу біологічної активності

В SAR/QSAR використовуються методи логічного висновку, на основі класичного, індуктивного, ймовірної, та інших видів логіки, методи аналізу повторень і кластеризації, методи лінійної і не параметричної регресії. Алгоритм прогнозу PASS відібраний серед досліджень, протягом десятків років. Для хімічних з'єднань С по його структурі, записаній в вигляді множини з m MNA-дескрипторів$\{D_1, \ldots, D_M\}$, оцінимо ймовірність $P(A_K\backslash C)$ того, що з'єднання С має активність A_K відповідно до формули Байеса:

$$P(A_K\backslash C)= P(C\backslash A_K)*P(A_K)/P(C)$$

Де $P(A_K\backslash C)$ – це ймовірність структури С при умові, що хімічне поєднання має активність A_K ; $P(A_K)$- апріорна ймовірність активності A_K ; P(C)- апріорна ймовірність структури С.

$$P(A_K\backslash C)=/ P(\underline{A_K}\backslash C)=[P(A_K)* P(A_K\backslash C)]/[P(A_K\backslash C)* P(\underline{A_K})]$$

Де $P(A_K\backslash C)$,$P(\underline{A_K}\backslash C)$,$P(\underline{A_K})$- відповідні ймовірності для $\underline{A_K}$ відсутності активності A_K .

Якщо припустити ,що дескриптори $D_1, \ldots, D_M$ незалежні від сукупності, то можна записати ймовірність $P(C\backslash A_K)$ і $P(C\backslash \underline{A_K})$ як похідна умовних ймовірностей для окремих дескрипторів:

$$P(C\backslash A_K)=P(D_1, \ldots, D_M\backslash A_K)=\Pi_i P(D_i\backslash A_K)$$

$P(C\backslash \underline{A}_K)=P(D_1, \ldots, D_M\backslash A_K)=\Pi_i P(D_i\backslash \underline{A}_K)$

Ці співвідношення приближені, тому MNA-дескриптори залежні від способу їх будови. Підставлення такого приближення дає наступний вираз:

$P(A_K\backslash C)/ P(\underline{A}_K\backslash C)=[P(A_K)/ P(\underline{A}_K)]* \Pi_i P(D_i\backslash A_K)/ \Pi_i P(D_i\backslash \underline{A}_K)$

При цьому, якщо активність не залежить від даного дескриптора, то $P(A_K\backslash D_i)=(A_K)$ а дескриптор не впливає на результат, а його вклад наближений до нуля. Це і є класичний результат ймовірного підходу, це результат має добре відомий недолік: вклад деяких дескрипторів, для яких умовна ймовірність активності наближена до 0 або 1, починає прямувати до безкінечності, і починає пригнічувати інші значення. Це сильно проявляється, коли для ймовірності $P(A_K\backslash D_i)$ часткові оцінки по результатам аналізу навчальної вибірки, а значення 0 і 1 швидше правило, ніж виняток.

Існує багато підходів для уникнення таких недоліків, і вони були випробовані в ході розвитку PASS. Кращий результат дало застосування Ln[p/(1-p)], так зване перетворення Фішера arcsin(2p-1) майже на всіх інтервалах змінної p її форма збігається. Точність прогнозу також збільшується після заміни суми вкладів дескрипторів на їх середнє значення, це компенсує припущення про незалежність дескрипторів. Описаний підхід пояснює чому PASS базується на специфікації B-статистики: по структурі молекул хімічних сполучень, записаних в вигляді множини з m MNA-дескрипторів $\{D_1, \ldots, D_M\}$, для кожної активності AK підраховують величини BK [6]

$B_K=(S_K-S_{0K})/(1-S_K*S_{0K})$

$S_K=\sin[\arcsin(2p(A_K/D_I)/m]$

$S_{0K}=2p(A_K)-1$

При цьому для кожного виду активності, для всіх видів дескрипторів $P(A_K\backslash D_I=1$, то $B_K=1$. Якщо для всіх дескрипторів $P(A_K\backslash D_I)=0$, то $B_K=-1$. Якщо зв'язки між дескрипторами і активністю A_K немає і $P(A_K\backslash D_I$ наближене до $P(A_K)$, то B_K наближено до 0.

До версії 1,709 алгоритм прогнозу PASS використовував наступну інформацію про “ структуру - активність”.

N- загальна кількість речовин в SAB base

N_I- кількість речовин, які містять дескриптори D_I в описаній структурі

N_K- кількість речовини, яка містить інформацію про активність в спектрі A_K

N_{iK}- кількість речовини, яка містить дескриптор D_I в описі структури, і активність в спектрі A_K

При цих даних вираховують часткові оцінки ймовірності $P(A_K)$ та $P(A_K \backslash D_I)$

$P(A_K)= N_K/N$, $P(A_K \backslash D_I)= N_{Ik/} N_I$

В PASS версії 1.703 і наступних версії оцінювання ймовірності $P(A_K)$ і $P(A_K \backslash D_I)$ вираховують в вигляді наступної суми по всім N речовин в SAB base:

$P(A_K \backslash D_I)=\sum_n f_n(A_K) g_n(D_I)/)=\sum_n g_n(D_K)$

$P(A_K)= \sum_n f_n(A_K))\sum_n g_n(D_I)/\sum_n \sum_i g_n(D_I)$

Де $f_n(A_K)$ і $g_n(D_I)$- це характеристичні функції речовини з номерами n k множини речовин, з активністю A_K в спектрі і дескриптор D_I в описаній структурі. $f_n(A_K)$ приймають значення 0 і 1, а $g_n(D_I)$- 0 і $1/m_n$, де m_n - число декрипторів молекули n і $\sum_i g_n(D_i)=1$.

Така модифікації алгоритму PASS не тільки дозволить повисити точність алгоритму прогнозування, а і відкриває нові можливості. Наприклад функцію $f_n(A_K)$ можна розглянути як міру належності до нечіткої множини речовин. Точно так само розглядають і вага дескрипторів $g_n(D_I)$, тоді дескриптори можуть бути любої однорідної природи. На цій основі й розроблявся метод кількісного прогнозу біологічної активності, і результати показують перевагу того способу на основі MNA-дескрипторів по порівнянню з 3D COMFA і COMSIA (трьох мірний порівняльний

аналіз молекулярних полів і порівнянь молекулярних подібностей). Головним призначення системи PASS полягає в прогнозі спектрів активності нових, ще не вивчених речовин. Тому загальний принцип алгоритму прогнозування, являється винятком з SAR base речовин з структурою молекул, еквівалентної структурі молекул речовини в спектрі біологічної активності. Якщо в SAR base знайдена еквівалентна структура з номером n, то цю речовину виключають з додавання. Середнє значення вкладів дескрипторів виконується по MNA-дескрипторам прогнозованої речовини. Для отримання якісного підходу, потрібно визначити крайні значення B- статистики, для кожного виду активності. З допомогою методів прийняття статистичних рішень можна, мінімізувати ризики. Але неможливо на період визначити всі види активності для всіх можливих практичних задач. Тому результати прогнозу спектра біологічної активності представлені у вигляді впорядкованого списку назв відповідних активностей і ймовірностей. Впорядковування виконують по зменшення відмінностей P_A-P_I, так як більш ймовірні види активності знаходяться в початку спектру. Прогнозований спектр активності можна аналізувати різними способами, але за замовчуванням в нього включають активності, для кожного $P_A \geq P_I$. [7]

1.5 Процедура навчання

Оцінки точності прогнозу PASS і залежності, потрібні для отримання ймовірності P_a і P_i по значенням B-статистики, являється кінцевим результатом процедури навчання, яка відбудеться в майбутньому. По даним SAR base, сформованим на основі навчальної вибірки, для кожної активності A_K для кожного з N_K активним і для кожного N-N_k неактивних речовин значень B статистики. Обрахунки проводяться в режимі ковзаючого контролю з виключенням одного, після виключення одного з'єднання з суми SAR base, для чого достатньо не включати його в суми.

Ймовірності P_a і P_i являються також по, будові, оцінкам ймовірності помилки 1-го і 2 – роду, відповідно. Їх можна розглядати і як засоби належності прогнозованої

речовини до нечітким множинам активних і не активних речовин. Всі ці інтерпретації ймовірностей еквівалентні і корисні для аналізу результатів прогнозу. На їх основі можна скорегувати прогнози, відповідне до рішення поставленої задачі.

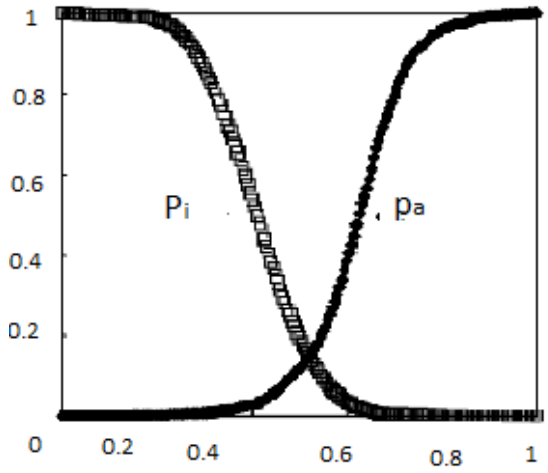


Рис. 1.1 - оцінка ймовірності $P_a(B)$ I $P_i(B)$

На рисунку 1 приведений приклад ймовірності $P_a(B)$ I $P_i(B)$ як функцій значень B-статистики для активності в SAR base версії 1.917.

Точка перенесення функції P_a і P_i відповідають рівності ймовірності помилки 1-го і 2-го роду, а значення в точці MEP= P_a = P_i являються оцінкою максимальної помилки прогнозу- характеристикою точності прогнозу конкретної біологічної активності. Загальна точність прогнозу оцінюється в PASS як середнє по всім прогнозованим видам активності значення MEP.[8]

Важливою особливістю алгоритму прогнозування PASS є стійкість до неповноти інформації про структури і спектри біологічної активності хімічних сполучень в навчальній вибірці.

1.6 Інтерпретація результатів прогнозування PASS

Важливо пам'ятати що ймовірність P_a відображає схожість молекулярної структури даної речовини з структурою молекул більш типових активних речовин в навчальній вибірці. Тому прямої кореляції величин P_a з кількісними

характеристиками активності, як правило немає. Не типова для вибірки речовина з нетиповою структурою молекул може мати низьке значення P_a навіть можливе $P_a \leq P_i$. це очевидно з способу побудови функцій $P_i(B)$ і $P_a(B)$, значення величини P_a для активних і P_i для неактивних речовин з навчальної вибірки розміщені строго рівномірно. Тут випливає і інтерпретація результатів прогнозу.

Якщо величина P_a дорівнює 0,- то для 90% активних речовин з вибірки значення B-статистики менше, ніж для досліджуваної речовини, і тільки для 10% буде більшим. Це також означає, що якщо ми відхилимо пропозицію , що речовина має активність, то в ми з ймовірністю 90% помилимось. Якщо величина P_a менша 0,5, а $P_a \geq P_i$, то більше половини активних речовин з вибірки мають значення B-статистики більше, ніж в даної речовини, якщо ми відхилимо пропозицію , що речовина має активність, то ми з ймовірністю 0,5% помилимось. В такому випадку ймовірність виявити даний вид активності експериментально невеликий, але дана структура буде оригінальної з шансом 50%. Обширно прогнозований спектр активності свідчить, що структура молекул даної речовини досить проста, і не зберігає ніяких особливих даних, для високої селективної біологічної роботи. А якщо при прогнозі виявилось, що в структурі є декілька нових MNA-дескрипторів, то структура мало схожа на любу відому з бази, а результати прогнозу можна розглядати як наближені.[9]

1.7 Використання прогнозу спектру біологічної активності в органічних сполуках

Використання PASS дозволяє вже на початкових етапах досліджень відібрати кандидатів, які можуть мати бажані види біологічної активності з малою ймовірністю викликати побічні ефекти. Використовувані в PASS MNA-дескриптори 1-го і 2-го рівнів не враховують ніяких просторових особливостей і охоплюють тільки малі локальні фрагменти молекул, в яких дальні атомі розділені не більше ніж на 4 зв'язка. Оскільки прогноз виконується по структурній формулі, він може бути отриманий ще на стадії планування синтезу.[10]

Часто лікарські препарати потрапляють в організм в формі хімічно модифікованих лікарських препаратів. Прогноз спектру біологічної активності для про ліки і їх активних речовин, показує, що в 74% випадках основний ефект ліків передбачається в PASS по структурній формулі його попередніх версій. Отримані результати прогнозу з допомогою PASS для 200 більш використовуваних лікарських препаратів , не просто збігається з 93% відомими фармакологічними ефектами і механізмами дій, а і вказують на нові можливі області застосування.

ВИСНОВОК ДО РОЗДІЛУ 1

Робота PASS заснована на аналізі залежностей «структура-активність» для речовин з навчальної вибірки, яка містить понад 45000 різноманітних біологічно активних речовин (субстанції відомих лікарських препаратів і фармакологічно активні сполуки). Навчальна вибірка постійно поповнюється новою інформацією про біологічно активних речовинах, що відбирається як з публікацій в науково-технічній літературі, так і з численних баз даних. Хімічна структура представлена в PASS у вигляді оригінальних MNA дескрипторів (Mulilevel Neighbourhoods of Atoms). MNA дескриптори мають універсальний характер і з досить гарною точністю описують різноманітні залежності «структура-властивість». Використовуваний в PASS математичний алгоритм був відібраний шляхом цілеспрямованого аналізу і порівняння ефективності для вирішення подібних завдань великого числа різних методів. Показано, що даний алгоритм забезпечує отримання стійких в статистичному сенсі залежностей "структура-активність" і, відповідно, результатів прогнозу. Це дуже важливо, оскільки включені в навчальну вибірку дані завжди мають певну неповноторність, як щодо охоплення всіх хімічних класів речовин, що мають конкретний вид активності, так і по відношенню до вивченості кожного окремого речовини на всі можливі види активності.

2 РОЗДІЛ

БІБЛІОТЕКИ ДЛЯ ПОБУДОВИ НЕЙРОМЕРЕЖ

2.1 Keras

Keras - це простий у використанні API для мови програмування Python, за допомогою якого дуже легко створювати моделі нейронних мереж. Він підходить для реалізації алгоритмів глибокого навчання і обробки природної мови. Модель нейронної мережі можна побудувати за допомогою всього декількох рядків коду.

Позитивні моменти Keras API

У Keras один з кращих прикладів документації. Вона являє кожну функцію послідовно і дуже докладно. Приклади коду також корисні і прості для розуміння.

У Keras є відмінна підтримка з боку співтовариства. Багато розробники вважають за краще Keras для участі в змаганнях з Data Science. Також багато дослідників публікують свій код і керівництва для широкої аудиторії.

Keras пропонує підтримку декількох бекенд-движків, включаючи Tensorflow, Theano і CNTK. Будь-який з них може бути обраний на основі вимог проекту.

Можна також тренувати модель Keras на основі одного движка, а перевіряти результати - на іншому. Поміняти движок в Keras також дуже легко. Для цього його ім'я потрібно просто записати в конфігураційному файлі.[11]

Keras дозволяє тренувати модель як на одному, так і на декількох GPU. Це забезпечує підтримку паралелізму даних і дозволяє обробляти великі обсяги. Ось деякі з недоліків інструменту.

1. Проблеми в низкорівневим API

Іноді виникають низькорівневі помилки бекенду. Це відбувається в тих випадках, коли робляться спроби виконати операції, для яких Keras не призначений.

Однак він не дозволяє змінювати що-небудь в бекенді. Отже, складно займатися налагодженням на основі балок з помилками.

2. Потрібні поліпшення деяких особливостей

Інструменти Keras для підготовки даних не такі хороші, як у випадку з іншими пакетами, наприклад scikit-learn. Вони не підходять для побудови базових алгоритмів машинного навчання: кластерного аналізу або методу головних компонент. Немає і можливості динамічного створення графіків.

3. Повільніший бекенд

Іноді Keras дуже повільний при роботі на GPU, а його операції займають більше часу в порівнянні з бекенд. Тому швидкістю доводиться жертвувати на догоду зручності використання.

2.2 Scikit-learn

Бібліотека Scikit-learn - найпоширеніший вибір для вирішення завдань класичного машинного навчання. Вона надає широкий вибір алгоритмів навчання з учителем і без вчителя. Одне з основних переваг бібліотеки полягає в тому, що вона працює на основі декількох поширених математичних бібліотек, і легко інтегрує їх один з одним. Scikit-learn широко використовується для промислових систем, в яких застосовуються алгоритми класичного машинного навчання, для досліджень.[12]

До завдань бібліотеки не входить завантаження, обробка, маніпуляція даними і їх візуалізація. З цими завданнями відмінно справляються бібліотеки Pandas і NumPy. Scikit-learn спеціалізується на алгоритмах машинного навчання для вирішення завдань навчання з учителем: класифікації (прогноз ознаки, безліч допустимих значень якого обмежена) і регресії (прогноз ознаки з речовими значеннями), а також для задач навчання без учителя: кластеризації (розбиття даних по класах , які модель визначить сама), зниження розмірності (подання даних в просторі меншої розмірності з мінімальними втратами корисної інформації) і детектування аномалій.

Бібліотека реалізує наступні основні методи:

Логістична регресія

Найчастіше використовується для вирішення задач класифікації (бінарної), але допускається і многоклассовая класифікація (так званий one-vs-all метод). Перевагою

цього алгоритму являєся те, що на виході для кожного об'єкта ми маємо вероятсность приналежність клас

Наївний Байес

Також є одним з найвідоміших алгоритмів машинного навчання, основним завданням якого є відновлення щільності розподілу даних навчальної вибірки. Найчастіше цей метод дає хорошу якість в задачах саме багатокласової класифікації,

До найближчих сусідів [13]

Метод kNN (k-Nearest Neighbors) часто використовується як складова частина більш складного алгоритму класифікації. Наприклад, його оцінку можна використовувати як ознака для об'єкта. А іноді, простий kNN на добре підібраних ознаках дає відмінну якість. При грамотній настройці параметрів (в основному - метрики) алгоритм дає часто гарна якість в задачах регресії

Дерева рішень

Classification and Regression Trees (CART) часто використовуються в задачах, в яких об'єкти мають категоріальні ознаки і використовується для задач регресії і класифікації. Дуже добре дерева підходять для многоклассовой класифікації

Метод опорних векторів

SVM (Support Vector Machines) є одним з найвідоміших алгоритмів машинного навчання, які застосовуються в основному для завдання класифікації. Також як і логістична регресія, SVM допускає многоклассовую класифікацію методом one-vs-all.

Крім алгоритмів класифікації і регресії, в Scikit-Learn є величезна кількість більш складних алгоритмів, в тому числі кластеризації, а також реалізовані техніки побудови композицій алгоритмів, в тому числі Bagging і Boosting.

Оптимізація параметрів алгоритму [14]

Одним з найскладніших етапів в побудові дійсно ефективних алгоритмів є вибір правильних параметрів. Зазвичай, це робиться легше з досвідом, але так чи інакше

доводиться робити перебір. На щастя, в Scikit-Learn вже є чимало реалізованих для цього функційGrid Search: метод для знаходження оптимальних гіперпараметров моделі шляхом побудови сітки з значень гіперпараметрів і послідовного навчання моделей з усіма можливими комбінаціями гіперпараметрів з сітки.

Це - лише базовий список. Крім цього, Scikit-learn містить функції для розрахунку значень метрик, вибору моделей, препроцесорінга даних та інші.

2.3. Pysmiles

Pysmiles – це програма для роботи з форматом SMILES. Молекули в ній зображені у вигляді графіків Networkx. Атоми - це вузли графіка, а зв’язки - ребра. Вузли можуть мати такі атрибути: [15]

- element: str. Описує елемент атома. За замовчуванням значення '*' означає невідоме.
- aromatic: bool. Чи є атом частиною (анти) -ароматичної системи. Значення за замовчуванням - помилкові.
- isotope: float. Маса атома. Значення за замовчуванням невідоме.
- hcount: int. Кількість неявних водню, приєднаних до цього атома. За замовчуванням до 0.
- charge: int. Заряд цього атома. За замовчуванням до 0.
- class: int. "Клас" цього атома. За замовчуванням до 0.

2.3.1. Читання SMILES

Функція read_smiles(smiles, explicit_hydrogen=False, zero_order_bonds=True, reinterpret_aromatic=True) може бути використана для розбору рядка SMILES. Він не повинен використовуватися для перевірки того, чи є рядок дійсним рядком SMILES --- функція робить дуже малу перевірку, чи має ваша струна SMILES хімічний сенс. Краї створеної молекули завжди матимуть атрибут "порядку". Вузли матимуть відповідні атрибути, поки вони визначені. Атоми, для яких елемент не відомий (*), не матимуть атрибута елемента. [16]

- explicit_hydrogen визначає, чи атоми водню повинні бути представлені як явні вузли у створеній молекулі, або неявно в атрибуті 'hcount'.

• zero_order_bonds визначає, чи повинні зв'язки нуля порядку (.) у рядку SMILES призводити до ребер у створеній молекулі.

• reinterpret_aromatic визначає, чи слід ароматичність переосмислювати та визначати з побудованої молекули, чи слід сприймати специфікації ароматичності з рядка SMILES (нижні регістри) як провідні. Якщо True, також встановить замовлення на облігації до 1 для зв’язків, які не входять до ароматичного кільця і мають порядок зв’язку 1,5. Якщо False, створить молекулу, використовуючи лише інформацію в рядку SMILES.

2.3.2. Запис SMILES

Функція write_smiles (molecule, default_element = '*', start = None) може використовуватися для запису рядків SMILES з молекули. Функція не перевіряє, чи має ваша молекула хімічний сенс. Натомість він пише SMILES-представлення молекули, яку ви надали, і нічого іншого. [17]

• default_element - це елемент, який потрібно використовувати для вузлів, у яких немає атрибута 'element'.

• start - це ключ від вузла, з якого слід починати перше проходження глибини. Щось розумне робиться, якщо не вказано.

2.3.3. Додаткові функції

Окрім цих двох основних функцій, піддаються ще чотири функції, які можуть допомогти у створенні хімічно відповідних молекул при мінімальній роботі.

• fill_valence (mol, respe_hcount = True, respe_bond_order = True, max_bond_order = 3) Ця функція заповнить валентність усіх атомів у вашій молекулі, збільшивши 'hcount' та, якщо зазначено, порядки зв’язків. Зверніть увагу, що він не використовує атрибут "зарядки", щоб знайти правильну валентність.

• repect_hcount: bool. Чи можна перезаписати наявні hcount.

• respect_bond_order: bool. Чи можна змінювати замовлення на облігації

• max_bond_order: int. Максимальний порядок облігацій, який буде встановлений.

• add_explicit_hydrogens (mol) Ця функція перетворює неявні водневі речовини, визначені атрибутами 'hcount', у явні вузли.

• remove_explicit_hydrogens (mol) Ця функція спрацьовує обернено add_explicit_hydrogens: вона видалить явні водневі вузли та добавить їх до відповідних атрибутів 'hcount'.

• correct_aromatic_rings(mol) Ця функція позначає всі (анти) -ароматичні атоми у вашій молекулі та встановлює всі зв'язки між (анти) -ароматичними атомами в порядку 1,5. Він заповнює валентність усіх атомів (див. Також fill_valence), перш ніж спробувати визначити, які атоми є ароматичними. Він працює, спочатку знаходячи всі атоми, які знаходяться в кільці. Далі, для кожного атома в кожному кільці перевіряється, чи атоми є гібридизованими sp2 (зауважте, що це невиразний термін. Строго кажучи, ми перевіряємо, чи є їх елементом щось ароматне, і чи мають вони 2 або 3 зв'язки.) . Нарешті, підраховується кількість електронів на кільце, і якщо це парне, атоми в кільці вважаються ароматичними. Ця функція є найбільш крихкою у всій бібліотеці, і я думаю, що вона дає неправильні відповіді в деяких випадках. Зокрема, для конденсованих (ароматичних) кільцевих систем (таких як індол) та кілець з екстрациклічними гетероатомами (O = C1C = CC = C1). [18]

ВИСНОВКИ ДО РОЗДІЛУ 2

Провівши аналіз існуючих рішень, можна вибрати кращі бібліотеки для нашого завдання. Keras- це зручна для використання бібліотека, для реалізації нейронних мереж. Ця бібліотека отримала добре описану документацію, в якій описані всі важливі функції. Також у Keras хороша репутація з сторони користувачів, а підтримка паралельної обробки даних на декількох GPU дозволяє обробити більші об’єми інформації, за коротший проміжок часу.

Також для опису структури і складу молекул хімічних елементів нам знадобиться Pysmiles. Ця програма дозволяє зручно візуалізувати будову елементів графічним способом, де атоми зображені в вигляді графів, а зв’язки це ребра. Також є можливість заповнити валентність всіх атомів у молекулі, а неявні водневі зв’язки перетворити у явні.

РОЗДІЛ 3

QSAR-задачі/моделювання

3.1 QSAR

В останній час термін QSAR отримує все більше і більше розповсюдження. Абревіатура QSAR перекладається з англійської як пошук кількісних співвідношень структура-властивість.

Одна з важливих задач сучасної хімічної науки полягає у встановленні залежності між структурами та властивостями речей (QSAR). Кількість нових синтезованих органічних з'єднань постійно збільшується, тому самої актуальної задачі є кількісне вгадування конкретних власних ресурсів для нових ще не синтезованих речей на підставі конкретних фізико-хімічних параметрів відокремлених з'єднань.

Історично все почалося з попитом вчених знайти кількість власних зв'язків між структурованими речовинами та їхніми властивостями та відтворити цю пам'ять у кількості власних зображень (у вигляді рівнянь). Це узагальнення повинно відобразити залежність одного набору цифр (властивостей) від іншого набору цифр (структур). Однак при цьому виникають складнощі. Відобразити цифрою властивість дійсно просто – фізіологічну активність серії речовин можна виміряти кількість. Проблеми в вираженні кількості структур хімічних сполук. Для такого рівняння в даний час в QSAR використовуються так звані дескриптори хімічної структури.

Дескриптор - параметр, що підтримує структуру організованих з'єднань, причому так, що підкреслюють деякі конкретні особливості структури. У принципі дескриптором може бути представлено будь-яке число, що можна розрізати з формулярної форми - молекулярну вагу, число визначених атомів, зв’язків або групи, молекулярний об'єм, часткові заряди на атомах. [19]

Для передбачуваної фізичної біологічної активності в QSAR зазвичай використовують наступні дескриптори: електронні ефекти (впливають на іонізацію або полярність з'єднань), стеричні особливості структури (відігрують важливу роль при оцінці міцності зв’язків досліджуваного з’єднання з молекулою-біомішенню), липофільність (можливість розчинятись в жирах характеризує здатність ліків

проходити крізь кліткові мембрани). Велику роль в QSAR відігрують топологічні дескриптори.

У методі QSAR структурна формула - граф, чисто математичне поняття (граф - математичний об'єкт, заданий множиною вершин і набором упорядкованих або невпорядкованих пар вершин (ребер)). Теорія графів дозволяє порахувати так звані інваріанти графів, які і розглядаються як дескриптори. Застосовуються також і складні фрагментні дескриптори, які оцінюють внесок різних частин молекули в загальну властивість. Вони значно полегшують дослідникам зворотне структурне конструювання невідомих сполук з потенційно високою активністю. Таким чином модель QSAR - це математичне рівняння (модель), за допомогою якого можна описати як фізіологічну активність (окремий випадок), так і взагалі будь-яку властивість, і цьому випадку правильніше говорити про QSPR - кількісному співвідношенні між структурою і властивістю. [20]

Методологія QSAR працює наступним чином. Спочатку групу сполук з відомою структурою і відомими значеннями фізіологічної активності (отриманими з експерименту) ділять на дві частини: тренувальний і тестовий набір. У цих наборах цифри, що характеризують активність, вже співвіднесені з конкретною структурою. Далі вибираються дескриптори (в даний час придумані багато сотень дескрипторів, проте реально корисних досить обмежене число; існують різні підходи до вибору найбільш оптимальних дескрипторів). На наступному етапі будують математичну залежність (підбирають математичне рівняння) активності від обраних дескрипторів для з'єднань з тренувального (навчального) набору і в підсумку отримують так зване QSAR-рівняння.

Правильність побудованого рівняння QSAR перевіряють на тестовому наборі структур. Спочатку обчислюють дескриптори для кожної структури з набору тестової вибірки, потім підставляють їх в QSAR-рівняння, розраховують значення активності і порівнюють їх з уже відомими експериментальними значеннями. Якщо для тестового набору спостерігається хороший збіг розрахункових і експериментальних значень, то дане QSAR-рівняння можна застосувати для передбачення властивостей нових, ще не синтезованих структур. Метод QSAR дозволяє, маючи в розпорядженні зовсім невелика кількість хімічних сполук з відомою активністю, передбачити

необхідну структуру (або вказати напрямки для модифікації) і тим самим різко обмежити коло пошуків.

У розвинених країнах роботи в області QSAR ведуться постійно зростаючими темпами оскільки застосування методів QSAR при створенні нових з'єднань із заданими властивостями дозволяє значно скоротити скринінг і здійснювати більш цілеспрямований синтез сполук, що володіють необхідним заданих комплексом властивостей. [21]

У методології QSAR виділяють пряму і обернену задачі. Пряма задача QSAR полягає в прогнозі активності на підстави знання структури. Зворотною завданням QSAR є конструювання хімічних структур із заданими величинами активностей.

Оскільки мова йде про біологічну активність, поняття QSAR близько поняття комп'ютерне моделювання лікарських препаратів (або комп'ютерний дизайн ліків), в англомовній літературі термін більш застоявся - Computer Aided Drug Design, або скорочено CADD - совокупність обчислювальних методів (в тому числі і методів QSAR) і програм, що використовуються для спрямованого молекулярного дизайну ліків (коректніше було б говорити про потенційні ліки або ж про з’єднання-лідерів). Хоча про чисто обчислювальному дизайні говорити ще рано, оскільки багато властивостей потенційних ліків в даний час можна визначити виключно експериментальним шляхом (обчислювальні ж оцінки носять якісний характер). CADD (комп'ютерний дизайн ліків) можна розглядати приватним (хоча і найбільш досліджуваним) напрямком CAMD (Computer Aided Molecular Design), комп'ютерного молекулярного дизайну. Таким чином комп'ютерний молекулярний дизайн являє собою совокупність підходів (методів, програм), які використовуються для молекулярного моделювання. Фактично експоненціальне розвиток обчислювальних методів в останнє десятиліття пов'язане головним чином з ростом потужностей обчислювальних ресурсів і в другу чергу - з розвитком різних методів і підходів.

3.2 Загальна постановка QSAR- задачі

Дамо визначення так званого QSAR / QSPR-аналізу (QSAR / QSPR - Quantitative Structure Analysis / Property Relationships, завдання «структура-властивість»). Він є найпоширенішим методом встановлення кількісних співвідношень між структурою та активністю сполук і являє собою статистичний підхід до проблеми. [22]

Мічений молекулярний граф G = {E, V} – позначений граф, вершини якого інтерпретуються як атоми молекули, а ребра – як валентні зв'язки між парами атомів. Мітки вершин і ребер (числа або символи) кодують атоми і зв'язку різної хімічної природи. В якості відміток можуть бути використані будь-які характеристики відповідних атомів (наприклад, тривимірні координати, символ хімічного елемента, заряд ядра, поляризованість, атомна вага, атомний радіус і ін.), а в якості міток ребер - будь-які характеристики відповідних зв'язків (кратність, довжини, порядки зв'язків, отримані з квантово-хімічних розрахунків, і т.д.).

Задача «структура-властивість»: Нехай задана навчальна (Або еталонна) вибірка - баз даних з N хімічних сполук, де:

1) i-е з'єднання представлено міченим молекулярним графом Gi, мають укладку в тривимірному просторі (тобто, для кожної вершини в якості міток задані її тривимірні координати);

2) або i-е з'єднання віднесено до Ci - одному з K класів активності (Наприклад, «активних», «слабоактивних», «неактивних» речовин) згідно досліджуваного властивості, або для нього задано чисельну значення досліджуваного властивості Ai.

Необхідно побудувати класифікуючу функцію F, яка одержує в якості аргументу довільний молекулярний граф з мітками того ж типу, і «Найкращим чином» відносить це з'єднання до одного з класів активності, або «найкращим чином» пророкує чисельне значення досліджуваного властивості. [23]

Яка з класифікують функцій «краще», дозволяє визначити функціонал якості φ (F). Наприклад, в якості функціоналу якості можна використовувати відсоток вірно класифікованих функцією F молекул з навчальної вибірки:

$$\varphi(F) = 1 - \frac{\sum_{i=0}^{n}(F(G_i) - A_i)^2}{\sum_{i=0}^{n} A_i^2}$$

Поставлену таким чином завдання пошуку класифікуючої функції будемо називати завданням «структура-властивість» або QSAR-завданням.

Дескриптором будемо називати будь-яку властивість, чисельне значення якого може бути обчислено для довільного молекулярного графа G.

Алфавітом дескрипторів будемо називати безліч всіх дескрипторів, які використовуються для аналізу навчальної вибірки, позначених різними символьними мітками. Визначення: Нехай алфавіт дескрипторів складається з M елементів. Вектором ознак молекулярного графа G будемо називати вектор

$\bar{x} = (x_i, \dots, x_M) \in R^M$, де xi- значення i-ого дескриптора, обчислене для G.

Матрицею «молекула-ознака» (матрицею ознак) для розглянутої навчальної вибірки будемо називати матрицю розміру N x M, в i-ому рядку якої варто вектор ознак i-ого з'єднання.

1.2.2. Основні етапи рішення QSAR-завдання.

У вищеописаних термінах завдання «структура-властивість» розбивається на дві частини:

1) етап опису:

Виходячи з формату молекулярних графів (типу відміток і ребер) вибирається алфавіт дескрипторів A. На основі цього алфавіту будується відображення з безлічі молекулярних графів в просторі ознак і формується матриця «молекула-ознака» для навчальної вибірки.

2) етап пошуку моделі функціональної залежності:

В результаті аналізу матриці «структура-властивість» на просторі ознак просторі будується модель функціональної залежності - класифікує функція F з найкращого прогностичної здатністю, тобто з найбільшим значенням функціоналу якості. [24]

3.3. Етап опису молекулярного графа.

Розглянемо два основних типи дескрипторів, на основі яких частіше всього будуються вектора ознак для молекулярних графів. Відразу відзначимо, що в даній роботі проводився аналіз матриць, побудованих переважно з дескрипторів другого типу. Їх докладний опис наведений в наступній главі.

Топологічні дескриптори:

Дані дескриптори будуються по топологічної матриці.

Визначення: топологічна матриця A = (aij) міченого молекулярного графа будується наступним чином: в ній елемент aii - це мітка i-ї вершини, aij - мітка ребра (i, j).

Очевидно, що при зміні нумерації вершин графа виходить, взагалі кажучи, інша матриця. Це є істотним недоліком опису хімічних структур в термінах матриць. Ця проблема привела до терміну «інваріант графа». [25]

Інваріант графа - це таке число (або функція, якщо мітка графа - символи), що обчислюється за матриці графа A, яке не залежить від нумерації вершин графа.

Молекулярний граф називається простим в наступному випадку:

- мітки вершин дорівнюють нулю
- якщо атоми пов'язані хімічним зв'язком, то їм зіставляється ребро з міткою «1», в іншому випадку мітка ребра дорівнює нулю, тобто простий граф відображає тільки наявність зв'язків між вершинами.

Топологічними індексами (ТІ) називають інваріанти простих графів. Часто це визначення переноситься і на інваріанти мічених графів, які можуть відображати не тільки топологію молекули, але і елементи електронного та просторової будови.

Топологічні індекси і широко використовуються як дескриптори при виконанні завдання «структура-властивість». Популярність даного підходу до опису молекулярної структури пов'язана з простотою і швидкістю обчислення ТІ, можливістю враховувати при їх побудові елементи електронного та просторової

будови, я також наявністю величезної кількості вдалих кореляцій виду «ТІ - властивість». Однак такий підхід має і очевидний недолік: він не дозволяє розрізняти різні конфігурації молекул і не враховує їх конформаційні особливості. [26]

Фрагментарні (структурні) дескриптори:

Використання даного типу дескрипторів засноване на виділенні в молекулах структурних фрагментів.

Структурним фрагментом молекулярного графа називається група вершин з заданими умовами на їх мітки або мітки їх зв'язків.

Після виділення фрагментів кожному фрагменту зіставляється структурний дескриптор, значення якого відповідає або наявності або відсутності даного фрагмента в молекулярному графі, або кількістю повторень фрагмента. У першому випадку отримуємо дескриптор, який приймає логічні значення, у другому - цілі невід'ємні. [28]

Структурні фрагменти і відповідні їм дескриптори поділяють на два типи:

1) 2D-дескриптори

В даному випадку не враховуються значення валентних кутів і евклідових відстаней між атомами, до уваги береться тривимірна структура фрагмента, важливі тільки зв'язку між атомами. Структурні 2D-фрагменти зазвичай мають 14 видів ланцюжків пов'язаних атомів з певними мітками вершин і ребер, котрі утворюють даний ланцюжок.

2) 3D-дескриптори

Дескриптори цього типу враховують тривимірну структуру фрагмента і зазвичай являють собою безліч вершин з заданими умовами на відстані між ними і на їх мітки. [29]

У загальному вигляді метод складається з наступних етапів:

1) Проводиться додаткова класифікація атомів (вершин молекулярного графа) на основі їх локальних властивостей (заряду, ексцентриситету вершини, яких-небудь

топологічних властивостей). Внаслідок цього мітка кожної вершини замінюється на іншу, що містить інформацію про локальні властивості.

2) В молекулах вибираються структурні фрагменти (атоми, ланцюжки пов'язаних атомів, групи атомів).

3) Кожному структурному фрагменту зіставляється символьне ім'я – тип фрагмента (наприклад, якщо розглядаються ланцюжка атомів, то «Ім'ям» ланцюжка може служити об'єднання символьних міток що входять в неї атомів).

4) Безліч фрагментів для всіх молекулярних графів вибірки об'єднуються.

5) Для кожного молекулярного графа і кожного фрагмента знаходимо значення відповідного структурного дескриптора (або кількість повторень, або наявність / відсутність в молекулярному графі).

У підсумку, отримуємо матрицю «молекула-ознака», що складається з «Структурних спектрів» молекул. [30]

Етап пошуку функціональної залежності:

Нагадаємо, що після формування матриці «структура-властивість» для навчальної вибірки необхідно побудувати класифікуючої функцію F (x1, x2, ..., xM), де (x1, x2, ..., xM) - вектор ознак молекулярного графа. причому, побудована функція повинна забезпечувати краще значення функціоналу якості.

Зазвичай вид класифікуючої функції F заздалегідь задається (наприклад, функція може бути лінійної, квадратичної та ін.) і залежить від ряду параметрів, які визначаються за навчальною вибіркою з'єднань. Найчастіше всього в якості F використовується лінійна функція. отримується рівняння називають лінійної регресійної моделлю.

Зауважимо, що отримана формулювання завдання на етапі пошуку функціональної залежності повністю збігається з формулюванням завдання розпізнавання. Тому для знаходження класифікуючої функції F можна використовувати будь-які методи розпізнавання і класифікації, описані в початку глави. [31]

Слід зазначити, що в завданні «структура-властивість» число дескрипторів M, як правило, значно перевищує число молекул в навчальній вибірці (M >> N), що ускладнює аналіз матриці «молекула-ознака». Для того щоб скоротити число дескрипторів, необхідно розглядати лише найбільш інформативні з них, тобто ті, які потенційно будуть значимі при побудові класифікуючої функції на просторі ознак. Це можна виконати як на етапі опису (наприклад, при еволюційному формуванні дескрипторів), так і на етапі аналізу матриці ознак. В даній роботі пропонується відбирати найбільш інформативні дескриптори шляхом взаємодії цих двох етапів - використання результатів етапу аналізу на етапі опису.

ВИСНОВКИ ДО РОЗДІЛУ 3

В останній час термін QSAR отримує все більше і більше розповсюдження. Абревіатура QSAR перекладається з англійської як пошук кількісних співвідношень структура-властивість. Основним завдання (завдання «структура-властивість») являє собою одну з задач області розпізнавання образів. Тому для її вирішення можуть бути застосовані всі методи вирішення завдань теорії розпізнавання.

Для передбачуваної фізичної біологічної активності в QSAR зазвичай використовують наступні дескриптори: електронні ефекти стеричні особливості структури, липофільність.

Методологія QSAR працює наступним чином. Групи сполук поділяють на дві частини: набір для тестів і для тренування. Велику роль в QSAR відігрують топологічні дескриптори, параметри які підтримують структурні з'єднання.

4 РОЗДІЛ
СТВОРЕННЯ ТА ТЕСТУВАННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Синдром набутого імунодефіциту (СНІД) є потенційно смертельним інфекційним захворюванням, спричиненим вірусом імунодефіциту людини (ВІЛ) . В даний час відсутність ефективних вакцин чи лікарських засобів досі залишається головною перешкодою у боротьбі з ВІЛ-інфекцією. Згідно з даними, повідомленими Всесвітньою організацією охорони здоров'я 17 травня 2017 року, за підрахунками, приблизно від 1,1 мільйона людей померли від захворювань, пов'язаних з ВІЛ у 2015 році, а 36,7 мільйона людей були заражені ВІЛ до кінця 2015 року. У 2016 році 18,2 мільйонам хворих на СНІД у всьому світі було призначено високоактивну антиретровірусну терапію (HAART), яка може різко знизити смертність ВІЛ-інфікованих пацієнтів шляхом інгібування реплікації ВІЛ .

Не нуклеозидні інгібітори зворотної транскриптази (NNRTI), як невід'ємна частина HAART, привернули широку увагу через потужну противірусну активність, високу специфічність та низьку цитотоксичність . NNRTI в основному інгібують зворотну транскриптазу (RT) ВІЛ 1-го типу (ВІЛ-1) шляхом зв'язування з гідрофобним кишенем, локалізованим приблизно 10 Å від каталітичного сайту ферменту . До цих пір повідомлялося про велику кількість NNRTI з різноманітними хімічними структурами, такі як дигідроалкоксибензилоксопіримідини, бензофенони, діарилові ефіри, діарилтріазини та діарилпіримідини (DAPY). Серед цих серій NNRTI, DAPY вважаються однією з найуспішніших сімей проти ВІЛ . Етравірин (TMC125) та рилпівірин (TMC278) (мал. 1), два представницьких члена DAPY, були затверджені відповідно американським управлінням харчових продуктів та лікарських препаратів відповідно. Порівняно з першим або другим поколінням NNRTI, таких як делавірдин та ефавіренц, етравірин та рилпівірин виявляють чудові потенції проти ВІЛ-1 дикого типу (WT) та стійких мутантів, таких як K103N. Однак реакції гіперчутливості, висип та синдром Стівенса – Джонсона спостерігалися у клінічних випадках етравірину . Крім того, у пацієнтів із СНІДом, у яких терапія

виявилася невдалою, було виявлено підвищену стійкість до препаратів до рильпівірину порівняно з ефавіренцем .

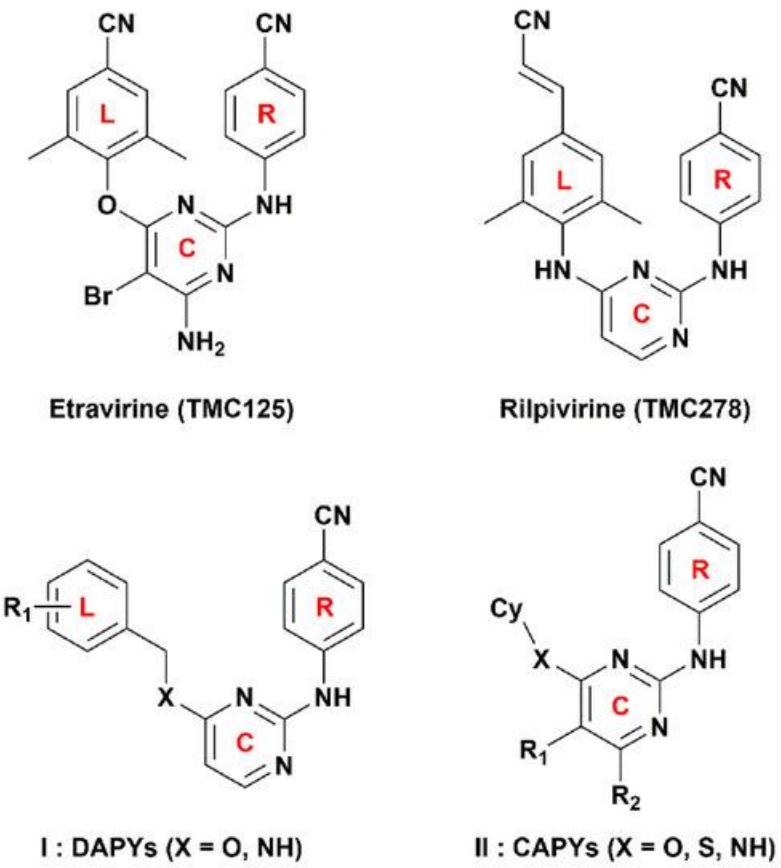


Рис.4.1 Структура діарилпіримідинів (DAPY) та циклоалкіл арилпіримідинів (CAPY).

Через те, що праве крило структури DAPY було підтверджено як ключовий фармакофор для активності проти ВІЛ-1 , структурні модифікації в основному здійснювались на лівому крилі, центральному піримідиновому кільці та їх лінкері (L – C міст) (рис. 1). Нещодавно ми синтезували серію DAPY (рис. 1), модифікуючи ліве крило, а також лінкер, і більшість з них виявляла потужну активність проти ВІЛ-1. Однак при модифікації лівого крила модифікації часто обмежувались різними заміщеними ароматичними групами без урахування циклоалкільних груп. Таким чином, у попередньому дослідженні ми також вивчали, чи введення циклоалкільної групи на лівому крилі DAPY для заміни фенільної групи може забезпечити новий структурний каркас для покращення активності проти ВІЛ-1. Кілька синтезованих циклоалкіл-арилпіримідинів (CAPY) також демонструють помірну активність проти

ВІЛ-1. Однак тривимірні кількісні зв'язки структура-активність (3D-QSAR) цих DAPY та CAPY та їх механізми взаємодії як NNRTI ВІЛ-1 були недостатньо вивчені. Для подальшого вивчення взаємозв'язку між інгібіторною активністю ВІЛ-1 ННРТІ та їх структурними особливостями в цій роботі були проведені дослідження 3D-QSAR, що включають порівняльний аналіз молекулярного поля (CoMFA) та порівняльний аналіз показників молекулярної подібності (CoMSIA) на серії DAPY і CAPY. Крім того, було проведено моделювання фармакофору та молекулярне стикування для дослідження структури зв'язування DAPY та CAPY з ферментом. Всі розроблені моделі могли б надати корисну інформацію про структурні модифікації при розробці нових та потужних DAPY як ННРТІ ВІЛ-1.

2. Результати та обговорення

2.1. CoMFA та статистичні результати CoMSIA

Класичні параметри моделей CoMFA та CoMSIA, включаючи значення q2, ONC, R2, r2pred, SEE та F, наведені на рис 4.2. Інші важливі параметри перевірки, такі як RMSE, MAE, r2, r20, r'20 , k, k ', r2m, r / 2m, Δr2m та r2m наведені на рис 4.3. Результати моделі CoMFA показали, що q2, R2, r2pred, MAE, RMSE, Δr2m та r2m були 0,679, 0,983, 0,884, 0,124, 0,160, 0,1215 (або 0,0026) і 0,7829 (або 0,9690) відповідно. Ці дані довели, що побудована модель CoMFA була надійною, а її точність прогнозування була прийнятною (r2pred> 0,5). Вклад стеричних та електростатичних полів становив відповідно 46.30% та 53.70%, що свідчить про важливий внесок електростатичних полів.

	Model	q^2	ONC	R^2	r^2_{pred}	SEE	F	Filed Contribution (%)				
								E	H	S	D	A
CoMFA	E + S	0.679	8	0.983	0.884	0.136	229.756	53.70	—	46.30	—	—
CoMSIA	E + H + S	0.721	9	0.972	0.734	0.180	114.912	52.40	32.40	15.20	—	—
	E + S + D + A	0.705	9	0.947	0.743	0.247	59.278	58.80	—	14.90	14.00	12.20
	E + H + D + A	0.695	11	0.987	0.826	0.125	198.216	45.10	34.00	—	12.60	8.20
	E + H + S + A	0.74	9	0.984	0.827	0.135	206.773	44.80	29.70	13.60	—	11.90
	E + H + S + D	0.703	9	0.972	0.698	0.178	117.713	48.70	29.70	14.30	7.30	—
	E + H + S + D + A	0.734	9	0.985	0.891	0.132	215.609	41.40	27.60	12.30	11.20	7.50

Рис4.2 Класичні статистичні параметри моделей порівняльного молекулярного поля (CoMFA)

Statistics	CoMFA (E + S)		CoMSIA (E + H + S + D + A)	
	Training Set	Test Set	Training Set	Test Set
RMSE	0.160		0.155	
MAE	0.124		0.108	
r^2	0.9834	0.8764	0.9848	0.8657
r_0^2	0.9830	0.8454	0.9848	0.8613
$r_0'^2$	0.9831	0.8750	0.9845	0.8144
k	1.0003	0.8408	1.0000	1.0848
k'	0.9831	0.9982	0.9997	0.9917
r_m^2	0.9637	0.7221	0.9848	0.8083
$r_m'^2$	0.9664	0.8436	0.9677	0.6696
Δr_m^2	0.0026	0.1215	0.0171	0.1387
$\overline{r_m^2}$	0.9690	0.7829	0.9763	0.7389

Рис 4.3 Зовнішні параметри перевірки моделей CoMFA та CoMSIA.

Для аналізу CoMSIA використовували різні комбінації полів дескрипторів для побудови різних моделей CoMSIA. Всі можливі комбінації полів були виконані для визначення оптимальної прогнозної моделі. Відповідно до експериментальних даних на рис 4.3, можна виявити, що модель, що складається з стеричних, електростатичних, гідрофобних, донороводородних зв'язків та акцепторних водень-зв'язкових полів, призвела до відносно більшого рівня q2, R2, r2pred та відносно нижчих показників SEE. Тому модель (S + E + H + D + A, таблиця 1) вважалася найкращою можливою комбінацією, яка присвоїла задовільні значення параметрам q2, R2, r2pred, MAE, RMSE, Δr2m та r2m, тобто 0,734, 0,985, 0,891, 0,108, 0,155, 0,1387 (або 0,0171) і 0,7389 (або 0,9763) відповідно. Відповідні внески стеричних, електростатичних, гідрофобних, акцепторних зв'язків водню та донорських полів складали 12,30%, 41,40%, 27,60%, 7,5% та 11,20% відповідно. Порівняно з моделлю CoMFA, здається, що модель CoMSIA демонструє дещо кращу прогнозованість. Було також встановлено, що електростатичні поля суттєво впливають на оптимальну модель CoMSIA. Ці результати свідчать про те, що побудовані моделі є потужними для прогнозування активності DAPY та CAPY. Польові внески показали, що електростатичні поля відіграють важливу роль у моделях CoMFA та CoMSIA.

Потім ми використовували моделі для прогнозування активності навчальних і тестових сполук. Фактичні та прогнозовані значення pEC50(−logEC50) DAPY та CAPY наведені на рис. 4.7. Кореляції між фактичними та прогнозованими значеннями pEC50 наведені на рисунку 2. Прогнозовані значення pEC50 були близькими до

експериментальних даних та більшості точок розташовувались на лінії тренда чи поблизу неї, що вказувало на прогнозованість та надійність обох моделей. Наведені вище результати свідчать про те, що побудовані моделі CoMFA та CoMSIA є розумними та мають можливість передбачати активність проти ВІЛ-1 навчальних та випробувальних сполук DAPY та CAPY.

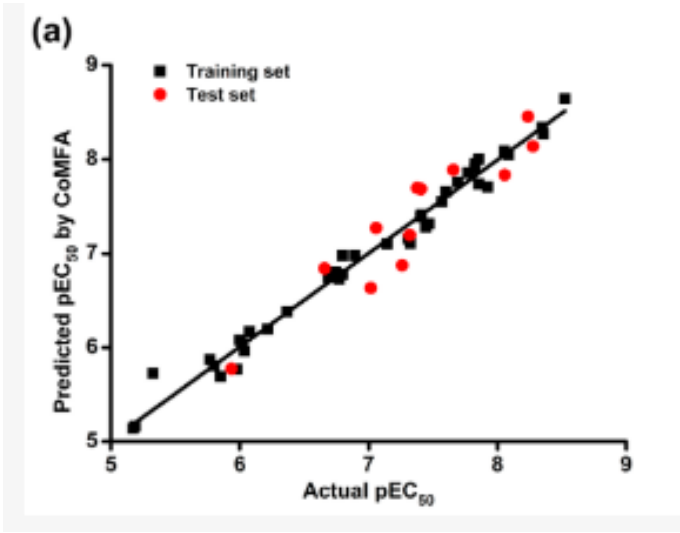


Рис 4.4 Діапазон розсіювання фактичних проти прогнозованих значень pEC50. (a) модель CoMFA;

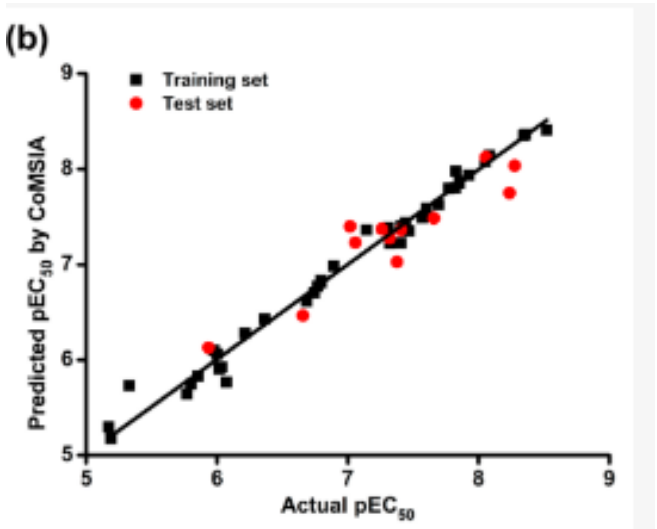


Рис 4.5 Діапазон розсіювання фактичних проти прогнозованих значень pEC50. (b) модель CoMSIA

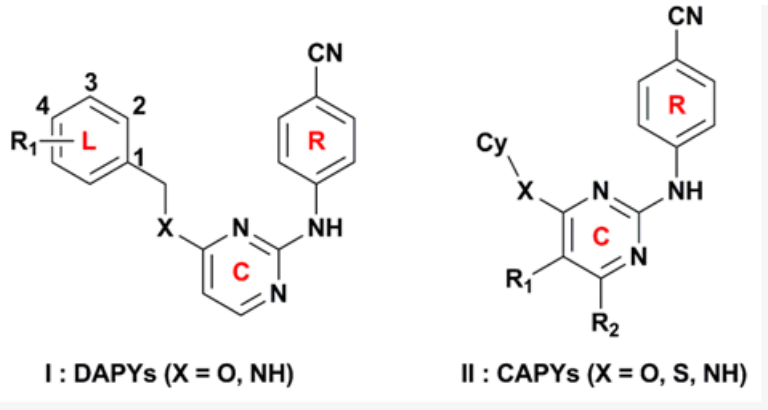


Рис4.6 Хімічні структури вибраних DAPY та CAPY

No.	Type	X	R_1	R_2	Cy	Actual pEC_{50}	CoMFA		CoMSIA	
							Predicted	Residuals	Predicted	Residuals
1	I	O	2-Cl	—	—	7.469	7.318	−0.151	7.355	−0.114
2	I	O	3-Cl	—	—	7.310	7.161	−0.149	7.384	0.074
3	I	O	4-Cl	—	—	6.076	6.172	0.096	5.769	−0.307
4	I	O	2-Br	—	—	7.444	7.283	−0.161	7.442	−0.002
5 *	I	O	3-Br	—	—	7.319	7.196	−0.123	7.277	−0.042
6	I	O	4-Br	—	—	5.328	5.727	0.399	5.733	0.405
7 *	I	O	2-F	—	—	8.237	8.454	0.217	7.750	−0.487
8	I	O	3-F	—	—	7.854	7.737	−0.117	7.848	−0.006
9	I	O	4-F	—	—	6.036	5.964	−0.072	5.927	−0.109
10	I	O	2-CH_3	—	—	6.796	6.974	0.178	6.831	0.035
11 *	I	O	3-CH_3	—	—	8.056	7.834	−0.222	8.124	0.068
12	I	O	4-CH_3	—	—	6.796	6.776	−0.020	6.811	0.015
13	I	O	2-OCH_3	—	—	7.569	7.551	−0.018	7.488	−0.081
14	I	O	3-OCH_3	—	—	7.602	7.654	0.052	7.587	−0.015
15	I	O	4-OCH_3	—	—	6.000	6.080	0.080	6.061	0.061
16	I	O	2-CF_3	—	—	5.770	5.870	0.100	5.644	−0.126
17	I	O	3-CF_3	—	—	6.215	6.197	−0.018	6.281	0.066

Рис 4.7 фактичні та прогнозовані значення pEC50

4.1 Контурні карти CoMFA та CoMSIA

Для візуалізації різних польових ефектів у тривимірних просторах, де модифікації могли б підвищити активність цільових сполук, контурні карти згодом створювались у моделях CoMFA та CoMSIA. З'єднання 43 з найбільшою активністю використовували в якості еталонної структури для ілюстрації всіх контурних карт. Стеричні та електростатичні контурні карти CoMFA та CoMSIA показані на малюнку 3. На стеричних полях зелені контури позначають стерично сприятливі об'ємні заступники, тоді як жовті контури вказують, де заступники є стерично несприятливими. Слід зазначити, що стеричні контурні карти CoMSIA подібні до

моделей CoMFA, що доводить узгодженість результатів. Як показано на малюнку 3, великий зелений контур, що оточує положення C3 (або C5) і C4 лівого фенільного кільця, вказує на те, що об'ємні групи тут сприятливі для посилення активності. Цей висновок може пояснити, чому активність сполуки 43 значно вище, ніж у інших сполук. Однак зв'язки структури та активності декількох сполук з однократною заміною в положенні C3 або C4 не збігалися з цим висновком, як видно, наприклад, для сполук 6(4-Br) <9 (4-F). Це може бути викликано іншими властивостями заміни або впливом інших полів і ще належить вивчити. З іншого боку, був невеликий жовтий контур, що оточує положення C2 лівого фенільного кільця, що дозволяє припустити, що об'ємні заступники на цій ділянці можуть бути несприятливими для активності, як це спостерігається для наступних сполук у цьому порядку: 4 (2 -Br) <1 (2-Cl) <7 (2-F) і 16 (2-CF3) <10 (2-CH3) <24 (2-H).

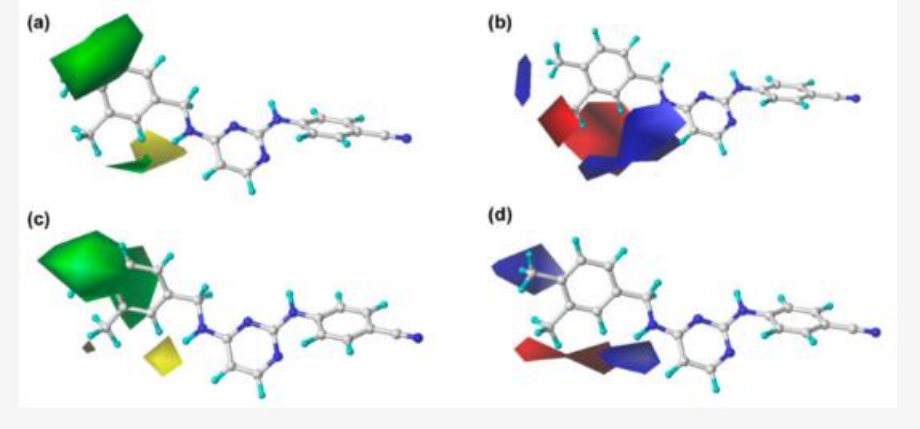


Рис 4.8 Контурні карти стеричних та електростатичних полів у поєднанні зі сполукою

Для електростатичних карт сині контури позначають ділянки, де позитивно заряджені заступники покращать інгібуючу активність, тоді як у червоних областях виявляються негативно заряджені заступники корисні для посилення активності. Як показано на малюнку 3, великий синій контур навколо положення C4 лівого фенільного кільця виявляє, що позитивно заряджені заступники в цьому положенні сприятливі для збільшення інгібіторної активності. Цей висновок може бути підтверджений такими прикладами: з'єднання 36 з метильним заступником у положенні C4 виявляло більш високу активність інгібування порівняно із сполуками

30 (4-Cl), 33 (4-Br) та 35 (4-F); порядок активності становив 36 (4-CH3)> 33 (4-Br)> 30 (4-Cl)> 35 (4-F). Більше того, великий синій нерегулярний контур біля лінкера між лівим крилом та центральним піримідином вказує на те, що присутність позитивно зарядженої групи в цій області сприятливо для біоактивності, що відповідає експериментальним даним, показуючи, наприклад, 27 (linker = –NH)> 24 (linker = –O). Дві невеликі червоні контури, що оточують положення C2 та C3 лівого фенільного кільця, відповідно, вказують на те, що наявність негативних зарядів у цих регіонах буде сприятливою для біоактивності. Наприклад, сполуки 1 (2-Cl), 4 (2-Br) і 7 (2-F) виявляли більшу активність, ніж з'єднання 10 (2-CH3), і порядок їх інгібіторної активності становив 7 (2-F))> 1 (2-Cl)> 4 (2-Br)> 10 (2-CH3). Цей результат відображається також у тому, що активність сполуки 34, що містить фтор (негативний заряд) у положенні C3, значно збільшується порівняно з активністю сполуки 41, що несе трифторметил (позитивний заряд).

Два великих помаранчеві контури розташовані поблизу положень C3 і C4 лівого фенільного кільця, що вказує на те, що гідрофобний заступник у цих двох положеннях буде сприятливим для інгібуючої активності. Наприклад, з'єднання 43 з метильними групами в цих двох зонах виявляло більш високу активність, ніж з'єднання 24 з гідрогенами. Крім того, існує білий контур, близький до положення C2 лівого фенільного кільця, що говорить про те, що присутність гідрофільної групи в цьому положенні може посилити інгібіторну активність. Наприклад, активність сполуки 7 з групою фтору в положенні C2 лівого фенільного кільця була вищою, ніж у з'єднання 24 з воднем у цьому положенні.

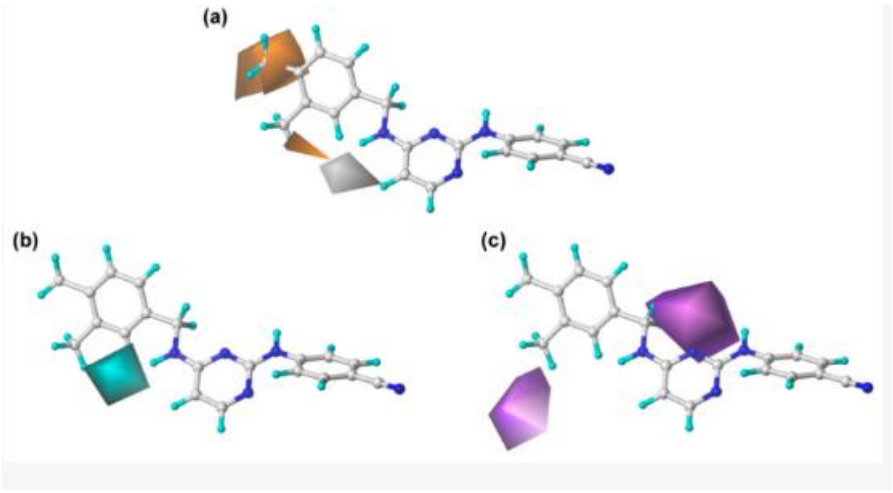


Рис 4.9 Контур CoMSIA відображається в поєднанні зі сполукою

Крім того, поля донорів та акцепторів водню також відіграють відносно важливу роль у біоактивності сполук. Як показано на малюнку 4b, c, блакитний контур оточує лінкер між центральним піримідином та лівим крилом, що означає, що групи донорів водневих зв'язків у цьому положенні будуть корисними для біоактивності. Це може бути підтверджено тим, що інгібуюча активність сполуки 27 (лінкер = -NH) була значно вищою, ніж активність сполуки 24 (лінкер = -O). Крім того, також можна помітити, що навколо положення C3 лівого фенільного кільця та X-заступника лінкера є два фіолетові контури (рис. 1) відповідно, що вказує на те, що акцепторні групи водневих зв'язків у цих положеннях не корисні для біологічної активності. Цей результат підтверджується біологічною активністю сполук, які містять атом кисню як заступника X, наприклад сполук 45, 46 та 47.

4.2 Фармакофорна модель

Фармакофорна модель була побудована з використанням дев'яти сполук з різноманітною структурою та відносно високою активністю як навчальний набір. Після генетичного алгоритму з лінійним призначенням гіпермолекулярної вирівнювання наборів даних (GALAHAD) було створено двадцять фармакофорних моделей, кожна з яких представляла різний компроміс серед конкуруючих критеріїв. Чим нижчі значення енергії деформації (SE), і чим вище значення стеричного перекриття (SO) та фармакофорної подібності (PhS), тим краще модель. За результатами експериментів було встановлено, що параметрами найкращої згенерованої моделі були: SE = 2.982, SO = 255.80 та PhS = 123.30. Модель фармакофору з вирівнюванням дев'яти сполук показана на малюнку 5, що свідчить

про задовільне накладення. Як показано на малюнку 5, пурпурова, зелена та блакитна сфери представляють атоми донора для зв'язків водню (DA), атоми акцепторних зв'язків водню (AA) та гідрофоби (HYs) відповідно. Найкраща модель складається з дев'яти особливостей фармакофору: двох DA-водневих зв’язків, чотирьох AA-водневих зв’язків та трьох центрів HY. Один з водневих зв’язків DAs - атом азоту імінової групи, а чотири AA водневих зв’язків відповідають атомам азоту піримідинового кільця, імінової групи та нітрильної групи відповідно. Ці особливості відображають важливість загального лісу DAPY / CAPY для інгібіторної активності. Інший водневий зв’язок DA розташований у атомі лінкера, що вказує, що групи донорів водневих зв'язків, такі як –NH в цьому положенні, можуть підвищувати інгібіторну активність, що відповідно до полів донора водневих зв’язків призводить до контурних карт CoMSIA . Три гідрофобні центри розташовані в центрі лівого фенільного кільця, у центрі піримідинового кільця та у центрі правого фенільного кільця відповідно, що говорить про те, що велика гідрофобна структура на лівому крилі сприятлива для активності. Ці результати узгоджуються з реальною діяльністю та стеричними полями контурних карт 3D-QSAR.

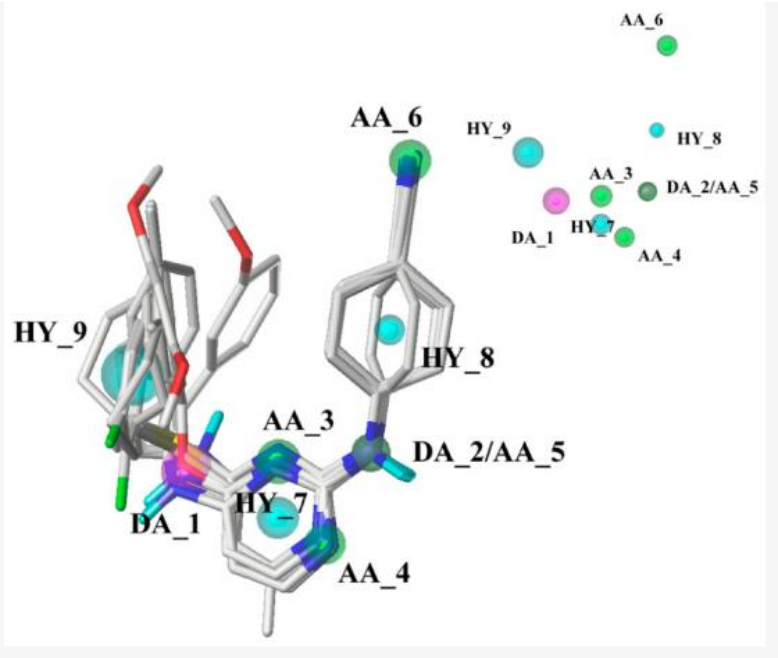


Рис 4.10 Фармакофорна модель з вирівнюванням дев'яти навчальних комплектів.

4.3 Молекулярний стикувальний аналіз

Для підтвердження надійності стикування, споріднений ліганд, тобто етравірин, який був витягнутий із кристалічної структури WT ВІЛ-1 RT (ID PDB: 3MEC), спочатку був повторно докрірований до місця зв'язування за допомогою док-рафа. Перероблену конформацію порівнювали з вихідною кристалографічною конформацією ліганду . Як показано на малюнку 6a, перероблений етравірин та кристалічний етравірин у комплексі майже повністю накладаються, а середньоквадратичне відхилення (RMSD) двох конформацій для всіх атомів становить 0,25 Å. Отримані результати свідчать про те, що метод док-серфінгу та використовувані параметри є розумними та надійними . Потім DAPY та CAPY були приєднані до місця зв'язування таким же чином. Сформований кишеню для зв’язування показано на малюнку 6b.

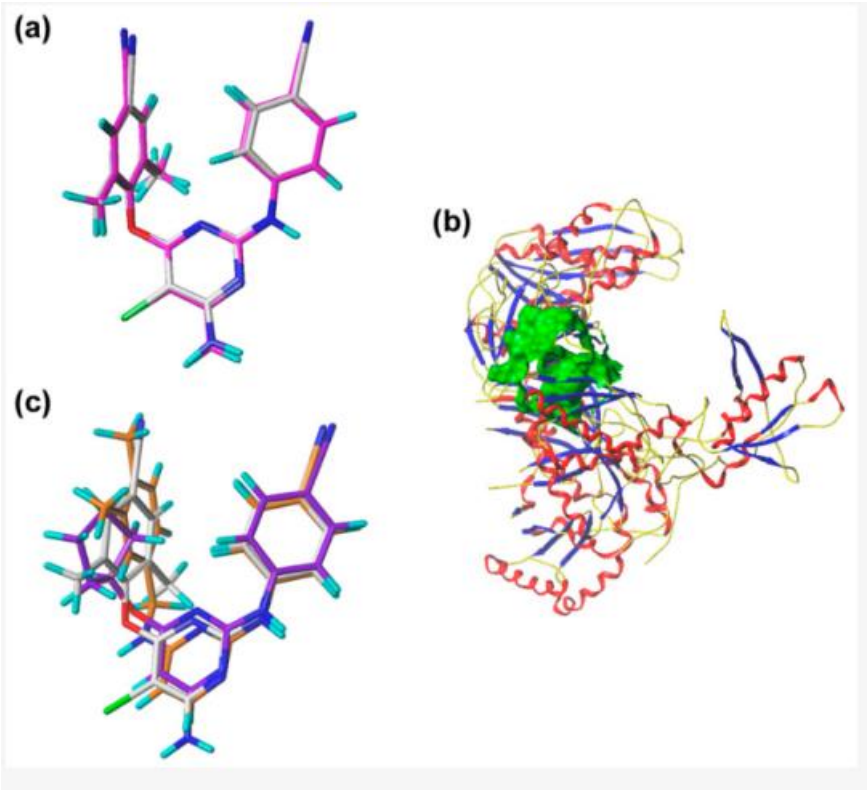
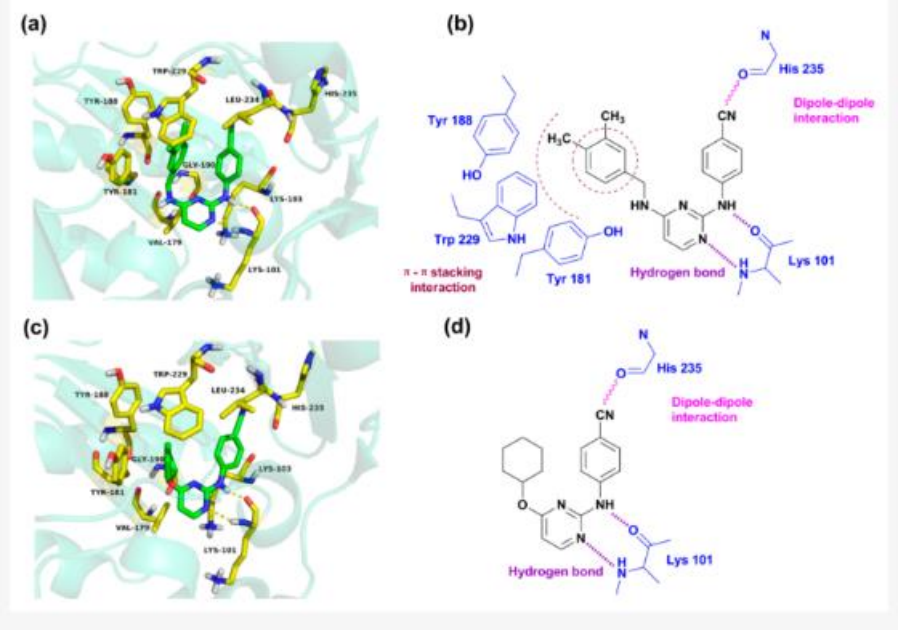


Рис 4.11 Накладення сполучених сполук та кишені для зв’язування. (a,b.c)

Після перевірки надійності стикування всі DAPY та CAPY були приєднані до кишені зв’язування. Накладення найактивнішої сполуки 43 та найменш активного з'єднання 46 з повторно складеним етравірином показано на малюнку 6c. Слід зазначити, що сполука 43 накладається з етравірином краще, ніж сполука 46, хоча їх док-формації мають аналогічну спрямованість. Док-станція з'єднання 43 (загальна

оцінка = 9,5247) була вище, ніж у з'єднання 46 (загальна оцінка = 8,2434), що відповідає їх діяльності.

На рис. 4.12 представлені детальні взаємодіючі способи сполук 43 і 46 в місці зв'язування ВІЛ-1 RT (3MEC). Як видно з малюнка 7a, c, ці дві сполуки мають однакову орієнтацію і приймають у кишені підкову або форму U-форми, як повідомлялося раніше. Цей результат узгоджується з нашим попереднім звітом, що залишок Lys101 може взаємодіяти з DAPY та CAPY через водневі зв’язки. Такі ж взаємодії спостерігалися і в режимі зв'язування сполуки 46.



Receptor	Ligand	Hydrogen-Bond Receptor	Hydrogen-Bond Donor	Distance (Å)	Angle (°)
3MEC	43	–O (Lys101 –C=O)	–N (–NH)	1.725	151.02
		–N (pyrimidine)	–N (Lys101 –NH_2)	2.302	168.45
	46	–O (Lys101 –C=O)	–N (–NH)	1.843	157.47
		–N (pyrimidine)	–N (Lys101 –NH_2)	2.477	163.30

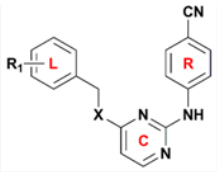
Рис 4.12 Результати стикування сполук

Було також встановлено, що деякі залишки амінокислот у кишені зв'язування, включаючи Tyr318, Tyr232, Phe 227, Trp239, Trp229, Pro225, Pro226, Met230, Ile94 та Val189, утворювали гідрофобні взаємодії із сполуками 43 та 46. Відповідно до фармакофорної моделі, можна також зробити висновок, що об'ємні ліпофільні заступники, такі як ароматичне кільце на лівому крилі DAPY, можуть здійснювати гідрофобні контакти з цими залишками амінокислот. Більше того, взаємодії ван дер Ваальса можуть бути встановлені між сполученими сполуками та амінокислотними залишками, такими як Leu100, Lys103, Val179, Gly190 та Leu234. Цианогрупа в

правому арильному крилі могла встановити диполь-дипольну взаємодію з карбонілом His235. Ці взаємодії можуть дозволити інгібіторам підтримувати підкову або конфігурацію у формі U.

Крім того, були знайдені π – π взаємодії укладання між лівим фенільним кільцем сполуки 43 та ароматичними залишками амінокислот, такими як Tyr188, Tyr181 та Trp229 . Як показано на рисунку 4.12(a), ліва фенільна група паралельна Tyr181 або Tyr188, а 4-CH3 на фенільному кільці вказує на дуже збережений Trp229. Однак взаємодії укладання π – π не знайдені в результатах стикування сполуки 46 через відсутність ароматичного кільця на лівому крилі. Результати показують, що циклогексильні або циклопентильні заступники на лівому крилі CAPY несприятливі для інгібіторної активності, що може бути наслідком втрати π – π взаємодії укладання.

На основі комбінованого аналізу результатів 3D-QSAR, фармакофору та молекулярного стикування були отримані взаємозв'язки структури та активності DAPY та згодом використані для проектування нових DAPY як потенційних HHPTI ВІЛ-1. Було розроблено десять нових DAPY, і їх діяльність проти ВІЛ-1 була передбачена CoMFA та найкращими моделями CoMSIA, як видно з Рис.4.13 У розробці цих нових DAPY було враховано декілька принципів. По-перше, ліве фенільне кільце було збережено як основний фрагмент у розроблених сполуках, оскільки він міг брати участь у взаємодії укладання π – π з залишками ароматичної амінокислоти у кишені зв’язування. По-друге, контурні карти полів донора водневих зв'язків та характеристики фармакофору вказують на те, що донори водневих зв'язків, розташовані у лівому лінкері, віддають перевагу для посилення активності, тому іміногрупа зберігається як лінкер замість атома кисню. По-третє, в ліву фенільну групу згідно аналізу контурних карт вводили різні заміни: (a) об'ємний, позитивно заряджений та / або гідрофобний заступник, такий як –CH2CH3, –CH (CH3) 2, –C (CH3) 3 і -NH2 у положенні C4; (b) негативно заряджену та / або гідрофобну групу, таку як –CN, –NO2 та –OOCCH3, у положенні C3; (c) невеликий, негативно заряджений та / або гідрофільний заступник, такий як –OH та –F, у положенні C2



No.	R_1	X	Predicted pEC_{50}	
			CoMFA	CoMSIA
53	4-CH_2CH_3	NH	8.655	9.508
54	4-$CH(CH_3)_2$	NH	8.765	10.212
55	4-$C(CH_3)_3$	NH	8.276	8.959
56	4-NH_2	NH	8.695	9.050
57	3-CN	NH	8.334	8.584
58	3-NO_2	NH	8.306	8.281
59	3-$OOCCH_3$	NH	8.227	8.873
60	3-OH	NH	9.020	9.026
61	2-OH	NH	7.845	8.690
62	2-F, 4-CH_3	NH	8.675	9.388

Рис 4.13 Хімічна структура нещодавно розроблених DAPY та їх прогнозовані значення pEC50 на основі моделей CoMFA та CoMSIA.

4.4 Матеріали та методи

Набір даних та вирівнювання

П'ятдесят дві сполуки DAPY та CAPY, що належать до класу NNRTI ВІЛ-1, були отримані в наших попередніх дослідженнях та використані для створення моделей 3D-QSAR . Всі сполуки були випадковим чином розділені на два набори, включаючи 40 сполук як навчальний набір для генерування моделі та 12 сполук як тестовий набір для перевірки моделі. Біоактивність усіх сполук перетворювалася на −logEC50 (pEC50). Їх структури та біоактивність наведені на рис 4.13. Усі розрахунки проводилися за допомогою програмного забезпечення SYBYL-X 2.1 (Tripos Inc., Сент-Луїс, штат Міссурі, США), що працює на робочій станції Windows 7. Енергетична оптимізація всіх молекул використовувала заряди Гастейгера-Гюккеля та силове поле Тріпоса методом градієнтного спуску, з критерієм конвергенції градієнта 0,005 ккал / моль · Å та максимальним коефіцієнтом ітерації 10000. Інші параметри були встановлені за замовчуванням. З метою отримання оптимального вирівнювання конформація з'єднання 43 з найбільшою біоактивністю була обрана як шаблон, а інші молекули були вирівняні на ній загальним вирівнюванням підструктури та ручним регулюванням. Загальний скелет (червоні атоми) для

молекулярного вирівнювання показаний на рис.4.14a, а накладені структури тренувального набору показані на малюнку рис. 4.14b.

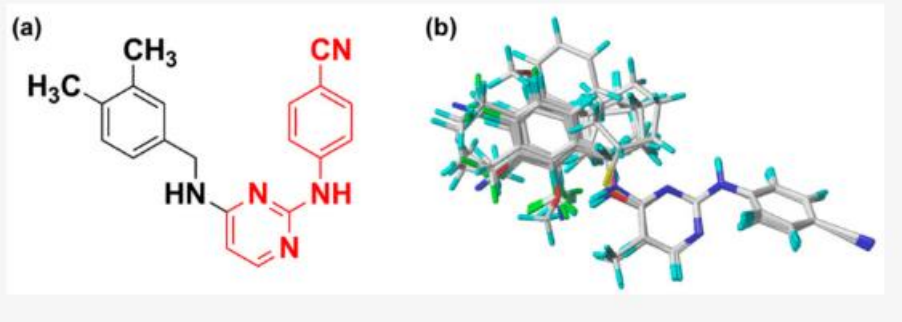


Рис 4.14 З'єднання 43 використовуване для вирівнювання всіх молекул

4.5 Моделі CoMFA та CoMSIA

Моделі 3D-QSAR були сформовані за допомогою методів CoMFA та CoMSIA, які могли б нам краще візуально зрозуміти зв’язок між структурними особливостями DAPY та їх інгібіторною активністю. Якість вирівнювання має важливий вплив на надійність та прогнозованість згенерованих моделей. Контурні карти моделей CoMFA та CoMSIA були графічно представлені, використовуючи тип поля «Stdev * Coeff». Для моделей CoMFA фізико-хімічні властивості, такі як стеричне та електростатичне поля, обчислювались у кожній точці сітки з регулярним проміжком сітки 2,0 Å, використовуючи гібридизований вуглецевий зонд sp3 з +1 зарядом. Для моделей CoMSIA п’ять фізико-хімічних властивостей, включаючи стеричні, електростатичні, гідрофобні, донорові зв'язки та акцепторні поля, були відповідно обчислені, використовуючи ту саму решітну коробку, яка також використовувалася в моделі CoMFA та вуглецевий зонд sp3 з +1 зарядом , +1 гідрофобність, +1 донор водородних зв'язків та +1 акцепторні властивості водневих зв'язків.

Для перевірки надійності створених моделей зазвичай проводять внутрішню і зовнішню перевірку. Для внутрішньої перевірки використовується метод "відхід один" (LOO) і витягує молекулу з набору даних як тестовий набір, а решту молекул розглядає як навчальний набір для створення моделей QSAR та прогнозування вилученої молекули. Ця методологія може отримати оптимальну кількість компонентів (ONC) на основі найвищого перехресного коефіцієнта кореляції (q2), який визначається наступним чином:

$$q^2 = 1 - \frac{\sum (r_{\text{pred}} - r_{\text{exp}})^2}{\sum (r_{\text{exp}} - r_{\text{mean}})^2}$$

де rpred, rexp і rmean представляють прогнозовані, експериментальні та середні значення pEC50 відповідно.

Для зовнішньої перевірки зазвичай використовується коефіцієнт кореляції прогнозування (r2pred) для оцінки якості прогнозування, який можна обчислити за наступною формулою:

$$r^2_{\text{pred}} = \frac{SDEP - PRESS}{SDEP}$$

де похибка стандартного відхилення прогнозування (SDEP) - це сума відхилень у квадраті між експериментальною діяльністю тестового набору та середньою активністю з'єднань навчального набору, а передбачувана сума залишків квадратів (PRESS) - це сума квадратичні відхилення між передбачуваною та експериментальною активністю для всіх сполук тестового набору.

Крім того, нова метрика r2m також використовується для додаткової внутрішньої перевірки (валідація навчального набору) та зовнішньої перевірки (перевірка набору тестів), представленої r2m (LOO) та r2m (тест) відповідно. Цей показник можна обчислити за наступним рівнянням:

$$r^2_{\text{m}} = r^2 \times (1 - \sqrt{r^2 - r^2_0})$$

де r2 і r20 - коефіцієнти визначення для найменшої регресії квадратів з і без перехоплення відповідно, що базується на передбачуваних значеннях pEC50 в осі x та експериментальних значеннях pEC50 в осі y.

Щоб уникнути завищення прогнозованої якості в результаті класичних метрик (q2 і r2pred), є кілька інших важливих параметрів перевірки, такі як RMSE (середньоквадратична помилка кореня), MAE (середня абсолютна помилка) , k, k ', і різні значення r2m (r / 2m, Δr2m та r2m) також використовуються для оцінки якості прогнозів. Модель може бути обрана для подальшого аналізу, якщо будуть задоволені наступні вимоги: q2> 0,5, R2> 0,6, r2pred> 0,5, 0,85 ≤ k (або k ') ≤ 1,15, r2 - r20 (r'20) r2 < 0,1, Δr2m <0,2 і r2m> 0,5 .

Молекулярне стикування

Добре відомо, що молекулярне стикування має велике значення для розуміння механізмів взаємодії між лігандом та рецепторним білком при проектуванні нових хімічних молекул. Для ефективного аналізу міжмолекулярних взаємодій між DAPY / CAPYs та ВІЛ-1 RT RT було здійснено молекулярне стикування за допомогою пакету док-серфінгу Sybyl-X 2.1. Кокристал WT ВІЛ-1 RT з TMC125 (3MEC), ліганд якого дуже схожий на DAPY, був отриманий від банку даних RCSB Protein. Перед стикуванням 3MEC готували шляхом видалення водних та сульфатних іонів та вилучення ліганду. Крім того, додавання водню та зарядів та обробка кінцевих залишків також проводили на 3MEC. Тоді "протомол" був сформований шляхом прийняття на основі ліганду режиму, і було сформовано відповідну зв'язуючу кишеню. Валідація надійності док-станції для серфлексу проводилася шляхом повторної переробки однорідного ліганду у кишеню, що зв'язує. Далі, всі DAPY були приєднані до кишені для зв’язування, і було отримано 20 можливих док-формацій з різними оцінками. Нарешті, доковані конформації найбільш активного з'єднання 43 та найменш активного з'єднання 46 були використані для аналізу механізму взаємодії.

ВИСНОВОК ДО РОЗДІЛУ 4

У цій роботі було проведено дослідження 3D-QSAR, фармакофору та докінгу на 52 похідних DAPY для дослідження взаємозв'язків між їх структурами та анти-ВІЛ-активністю. Були побудовані моделі CoMFA та CoMSIA з високою статистичною значимістю та хорошою прогностичною спроможністю та створена потенційна фармакофорна модель. Результати стикування продемонстрували режими взаємодії DAPY в кишені зв'язування ВІЛ-1 RT та припустили, що ліве фенільне кільце DAPY відіграє ключову роль у активності проти ВІЛ-1. Модель фармакофору та контурні карти 3D-QSAR дозволили візуалізувати вимоги до функцій для покращення активності. Було розроблено кілька нових DAPY з посиленою передбачуваною активністю. Однак ці новосформовані DAPY залишаються синтезувати та тестувати. Їх фармакокінетичні профілі також повинні бути визначені, якщо вони виявляють покращену інгібіторну активність проти ВІЛ-1 RT. В цілому побудовані моделі та отримана інформація можуть бути застосовані для подальшого раціонального проектування нових та потужних аналогів DAPY.

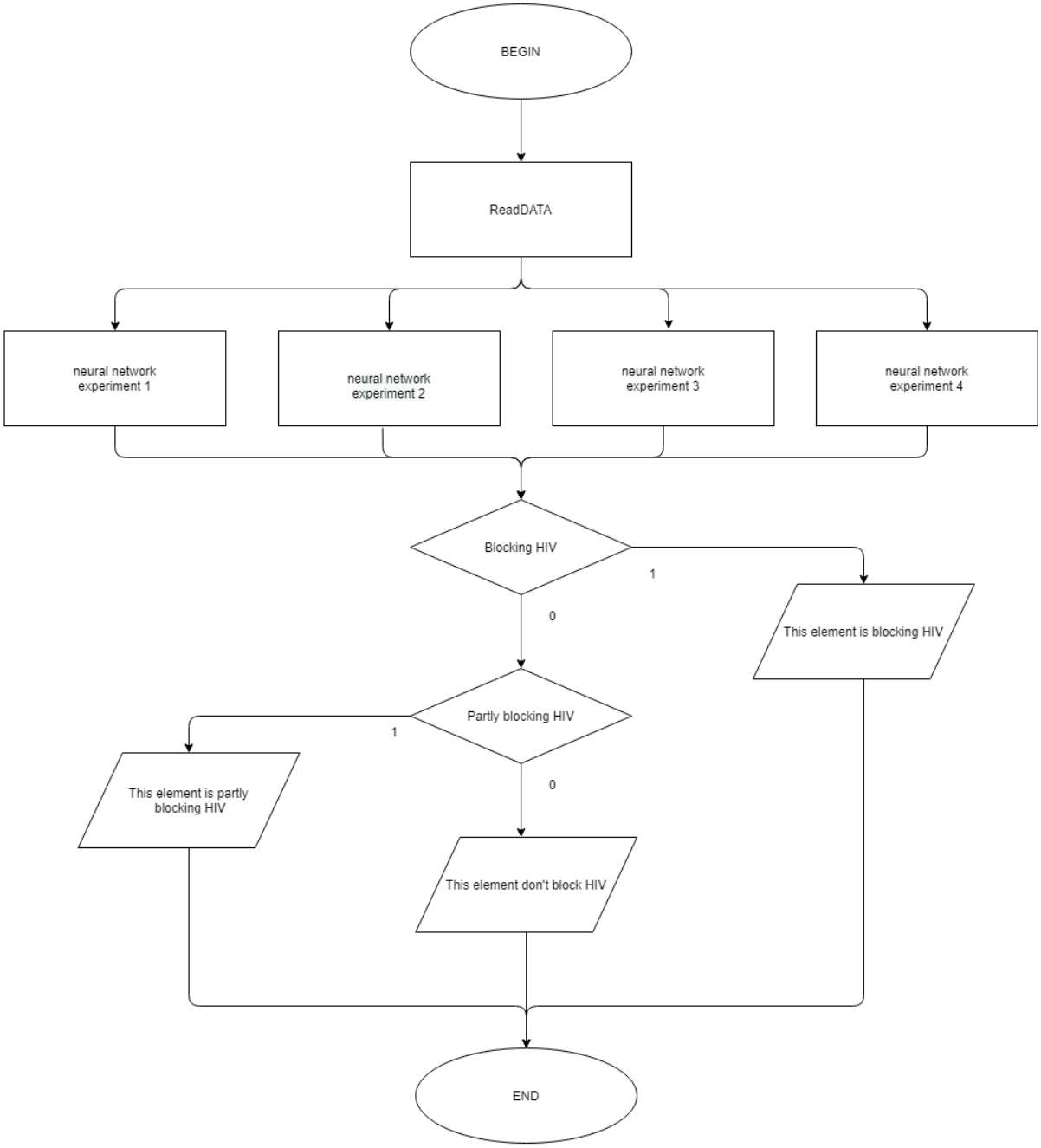
ВИСНОВКИ

В даний час в розвинених країнах пошук нових ліків переважно заснований на дослідженні величезних масивів хімічних речовин по відношенню до порівняно невеликого числа необхідних видів біологічної активності. Властивості виявлених таким шляхом базових структур в подальшому оптимізуються шляхом синтезу і дослідження великого числа їх аналогів. При цьому багато видів біологічної дії, якими вивчаються речовин, але є "побічними" по відношенню до обраних напрямом досліджень, залишаються невивченими. «У спадок» від СРСР Росії, Україні, Казахстану, Вірменії, Молдові та іншим незалежним державам дісталася досить розвинена органічна хімія синтетичних і природних речовин. Згідно з оцінками провідних фармацевтичних фірм, найбільші масиви хімічних сполук, синтезованих «вручну» (без використання комбінаторної хімії) і доступних в даний час для скринінгу, забезпечуються хіміками з країн СНД. Але при наявності досить об’ємну колекцію різноманітних хімічних сполук, країни СНД мають вкрай обмежені можливості для їх експериментального тестування, що вимагає ретельного відбору потенційно перспективних речовин вже на ранніх стадіях дослідження. Такий відбір може бути здійснений на основі комп'ютерного прогнозу спектру біологічної активності хімічних сполучень. Під спектром біологічної активності ми розуміємо всю сукупність фармакологічних ефектів, біохімічних механізмів дії і видів специфічної токсичності, які речовина може проявити при взаємодії з біологічними об'єктами. В рамках такого визначення ми абстрагуємося від багатьох чинників, що впливають на кількісні характеристики біологічної активності (об'єкт, доза, шлях введення і т.д.), і розглядаємо біологічну активність як «внутрішнє» властивість речовини, яке проявляється при відповідних умовах в експерименті або клініці

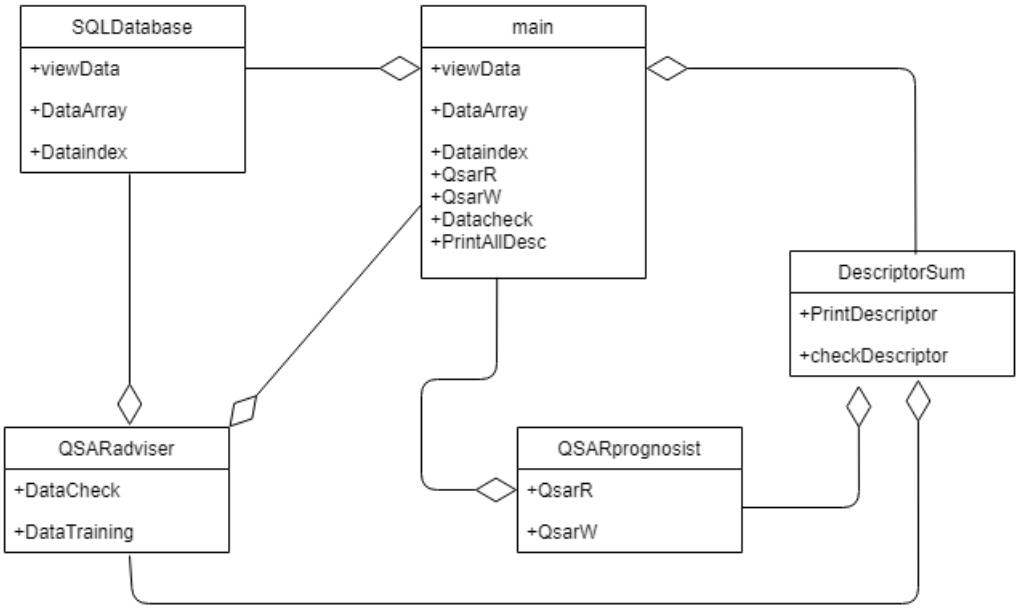
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Organic Chemical – Weinheil: Wiley VCH, 2001. – 363 с.
2. Баранбойм Г. М. Біологічно активні речовини / Г. М. Баранбойм. – москва: наука, 1986. – 375 с.
3. Арчаков А. И. Біоінформатика / А. И. Арчаков. – москва, 1999. – 47 с.
4. Livingsone D. Analysis for Chemist / Livingsone., 1995. – 654 с.
5. Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. Программная система. Практические применения. - М.: ФАЗИС, 2006.
6. Журавлев Ю. И. Избранные научные труды. - М.: Магистр, 1998.
7. Ryazanov V.V. Recognition Algorithms Based on Local Optimality Criteria. – Pattern Recognition and Image Analysis, 1994, vol. 4, no. 2.
8. Ryazanov V.V., Sen'ko O.V., Zhuravlev Yu.I. Methods of Recognition and Prediction Based on Voting Procedures. – Pattern Recognition and Image Analysis, 1999, vol. 9, no. 4.
9. Станкевич М.И., Станкевич И.В., Зефиров Н.С. Успехи химии, 57.
10. Lee, B. and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. / J. Mol. Biol. 1971, №55, pp.379-400.
11. опис бібліотеки Scikit [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/.
12. опис бібліотеки Scikit [Електронний ресурс] – Режим доступу до ресурсу: https://neurohive.io/ru/osnovy-data-science/vvedenie-v-scikit-learn/.
13. опис бібліотеки Scikit [Електронний ресурс] – Режим доступу до ресурсу: https://habr.com/ru/company/mlclass/blog/247751/
14. опис бібліотеки Scikit [Електронний ресурс] – Режим доступу до ресурсу: https://www.bigdataschool.ru/wiki/scikit-learn.
15. опис бібліотеки Pysmiles [Електронний ресурс] – Режим доступу до ресурсу: https://pypi.org/project/pysmiles/.
16. опис бібліотеки Pysmiles [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/pckroon/pysmiles.
17. опис бібліотеки Pysmiles [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/jhosmer/PySmile.
18. опис бібліотеки Pysmiles [Електронний ресурс] – Режим доступу до ресурсу: https://libraries.io/pypi/pysmiles.
19. Chauvin Y. Catalytic dimerization of alkenes by nickel complexes in organochloroaluminate molten salts / Y. Chauvin, B. Gilbert, I. Guibard // Chem. Comm. – 1990. – V. 23. – P. 1715-1716.
20. Fully plastic actuator through layer-by-layer casting with ionic liquid-based bucky gel / T. Fukushima, K. Asaka, A. Kosaka [et al.] // Angew. Chem. Int. Ed. – 2005. – V. 44. – P. 2410-2413.

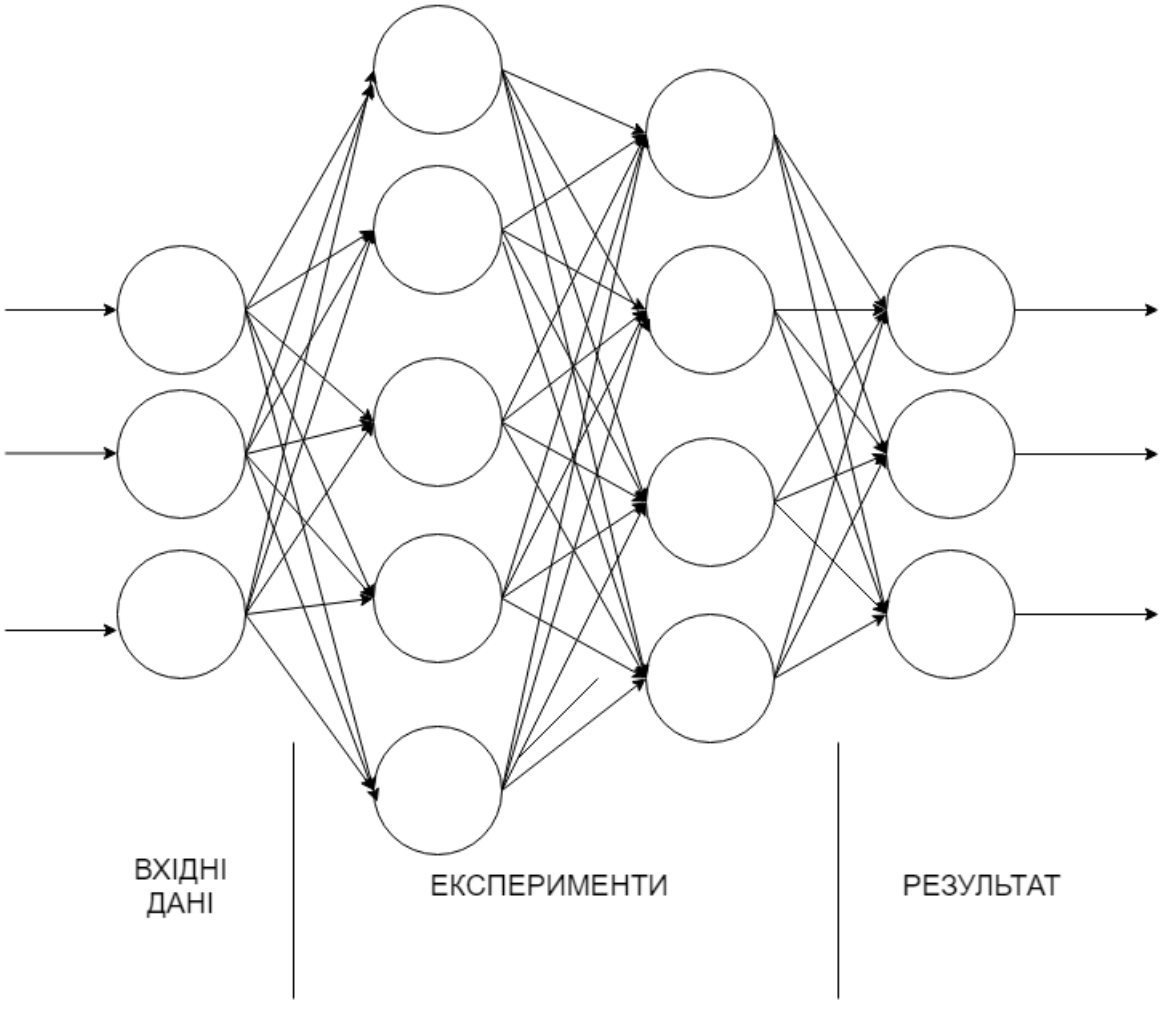
21. Protic ionic liquids as electrolytes for lithium-ion batteries / S. Mennea, J. Piresb, M. Anoutib [et al.] // Electrochemistry Communications. – 2013. - V. 31. – P. 39-41.
22. Chiappe C. Point-Functionalization of Ionic Liquids: An Overview of Synthesis and Applications / C. Chiappe, C.S. Pomelli // European Journal of Organic Chemistry. – 2014. – V. 28. – P. 6120-6139.
23. Enzymatic synthesis of caffeic acid phenethyl ester analogues in ionic liquid / A. Kurata, Y. Kitamura, S. Erie [et al.] // J. Boitechnol. – 2010. – V. 148, №23. – P. 133-138.
24. New ionic derivatives of betulinic acid as highly potent anti-cancer agents// C. Suresh, H. Zhao, A. Gumbs [et al.] // Bioorg. Med. Chem. Lett. – 2012. – V. 22, №4. – P. 1734-1738.
25. Pernak J. Synthesis and anti-microbial activities of choline-like quaternary ammonium chorides / J. Pernak, P. Chwala // Chem. Eur. J . – 2003. – V. 38. P. 1035-1042.
26. Лекарственные препараты: справочник Компендиум 2015 / Под ред. В.Н. Коваленко, А.П. Викторова // – К., 2015. 70
27. Assessing toxicity and biodegradation of novel, environmentally benign ionic liquids (1-alkoxymethyl-3-hydroxypyridinium chloride, saccharine and acesulfamates) on cellular and molecular level / M. Stasiewicz, E. Mulkiewicz, R. Tomczak-Wandzel [et al.] // Ecotox. Environ. Saf. – 2008. – V. 71. – P. 157-165.
28. Docherty K.M. Toxicity and antimicrobial activity of imidazolium and pyridinium ionic liquids / K.M. Docherty, C.F. Kulpa // Green Chem. – 2005. – V. 7. – P. 185-189.
29. Influence of ionic liquids on the growth of Escherichia coli / S.-M. Lee, W.-J. Chang, A.-R. Choi [et al] // Korean. J. Chem. Eng. – 2005. – V. 22. – P. 687690.
30. Ganske F. Growth of Escherichia coli, Pichia pastoris and Bacillus cereus in the presence of the ionic liquids [BMIM] [BF4] and [BMIM] [PF6] and organic solvents / F. Ganske, U.T. Bornscheuer // Biotechnol. Lett. – 2006. – V. 28. – P. 465-469.
31. Lewis M.A. Algae and vascular plant tests, In: Fundamentals of Aquatic Toxicology: Effects, Environment Fate, and Risk Assessment / M.A. Lewis // Taylors & Francis, Washington, DC, USA – 1995. – P. 135–170.



					ІАЛЦ.467800.004 А1		
Зм.	Арк.	№ докум.	Підпис	Дата			
Розробив		*Келеберда А.М.*			*Принципова схема алгоритму*	Літ.	Аркуш 1 / Аркушів 1
Перевірив		*Новотарський М.А.*					
Реценз.						*НТУУ «КПІ», ФІОТ, ІО-63*	
Н. Контр.		*Сімоненко В.П.*					
Затвердив							



					ІАЛЦ.467800.005 А2		
Зм.	Арк.	№ докум.	Підпис	Дата			
Розробив		*Келеберда А.М.*			*Функціональна схема*	Літ.	Аркуш 1 / Аркушів 1
Перевірив		*Новотарський М.А.*					
Реценз.						*НТУУ «КПІ», ФІОТ, ІО-63*	
Н. Контр.		*Сімоненко В.П.*					
Затвердив							



					ІАЛЦ.467800.006 А3		
Зм.	Арк.	№ докум.	Підпис	Дата			
Розробив		*Келеберда А.М.*			*Структурна схема*	Літ.	Аркуш 1 / Аркушів 1
Перевірив		*Новотарський М.А.*					
Реценз.						*НТУУ «КПІ», ФІОТ, ІО-63*	
Н. Контр.		*Сімоненко В.П.*					
Затвердив							